Minho Lee Akira Hirose Zeng-Guang Hou
Rhee Man Kil (Eds.)

LNCS 8228

Neural Information Processing

20th International Conference, ICONIP 2013
Daegu, Korea, November 2013
Proceedings, Part III

3 Part III

Springer

Lecture Notes in Computer Science 8228

Minho Lee Akira Hirose Zeng-Guang Hou
Rhee Man Kil (Eds.)

Neural Information Processing

20th International Conference, ICONIP 2013
Daegu, Korea, November 3-7, 2013
Proceedings, Part III

Springer

Volume Editors

Minho Lee
Kyungpook National University
1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, Korea
E-mail: mholee@knu.ac.kr

Akira Hirose
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
E-mail: ahirose@ee.t.u-tokyo.ac.jp

Zeng-Guang Hou
Chinese Academy of Sciences, Institute of Automation
Key Laboratory of Complex Systems and Intelligence Science
Beijing 100190, China
E-mail: zengguang.hou@ia.ac.cn

Rhee Man Kil
Sungkyunkwan University
2066, Seobu-ro, Jangan-gu, Suwon 440-746, Korea
E-mail: rmkil@skku.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-42050-4 e-ISBN 978-3-642-42051-1
DOI 10.1007/978-3-642-42051-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013951838

CR Subject Classification (1998): I.5, I.4, I.2, F.1, H.3, H.4, F.2, G.3, J.3, C.3

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume is part of the three-volume proceedings of the 20th International Conference on Neural Information Processing (ICONIP 2013), which was held in Daegu, Korea, during November 3–7, 2013. ICONIP is the annual conference of the Asia Pacific Neural Network Assembly (APNNA). This series of conferences has been held annually since ICONIP 1994 in Seoul and has become one of the premier international conferences in the areas of neural networks.

Over the past few decades, the neural information processing community has witnessed tremendous efforts and developments from all aspects of neural information processing research. These include theoretical foundations, architectures and network organizations, modeling and simulation, empirical study, as well as a wide range of applications across different domains. Recent developments in science and technology, including neuroscience, computer science, cognitive science, nano-technologies, and engineering design, among others, have provided significant new understandings and technological solutions to move neural information processing research toward the development of complex, large-scale, and networked brain-like intelligent systems. This long-term goal can only be achieved with continuous efforts from the community to seriously investigate different issues of the neural information processing and related fields. To this end, ICONIP 2013 provided a powerful platform for the community to share their latest research results, to discuss critical future research directions, to stimulate innovative research ideas, as well as to facilitate multidisciplinary collaborations worldwide.

ICONIP 2013 received tremendous submissions authored by scholars coming from 30 countries and regions across six continents. Based on a rigorous peer review process, where each submission was evaluated by at least two qualified reviewers, about 270 high-quality papers were selected for publication in the prestigious series of *Lecture Notes in Computer Science*. These papers cover all major topics of theoretical research, empirical study, and applications of neural information processing research.

In addition to the contributed papers, the ICONIP 2013 technical program included a keynote speech by Shun-Ichi Amari (RIKEN Brain Science Institute, Japan), 5 plenary speeches by Yoshua Bengio (University of Montreal, Canada), Kunihiko Fukushima (Fuzzy Logic Systems Institute, Fukuoka, Japan), Soo-Young Lee (Brain Science Research Center, KAIST, Korea), Naftali Tishby (The Hebrew University, Jerusalem, Israel) and Zongben Xu (Xi'an Jiatong University, China). This conference also featured invited presentations, regular sessions with oral and poster presentations, and special sessions and tutorials on topics of current interest.

Our conference would not have been successful without the generous patronage of our sponsors. We are most grateful to our sponsors Korean Brain Research

Institute, Qualcomm Korea. We would also like to express our sincere thanks to the International Neural Network Society, European Neural Network Society, Japanese Neural Network Society, Brain Engineering Society of Korea, and The Korean Society for Cognitive Science for technical sponsorship.

We would also like to sincerely thank honorary chair Shun-ichi Amari, Soo-Young Lee, the members of the Advisory Committee, the APNNA Governing Board and past presidents for their guidance, the organizing chair Hyeyoung Park, the members of the Organizing Committee, special sessions chairs, Publication Committee and publicity chairs, for all their great efforts and time in organizing such an event. We would also like to take this opportunity to express our deepest gratitude to the members of the Program Committee and all reviewers for their professional review of the papers. Their expertise guaranteed the high quality of the technical program of the ICONIP 2013!

Furthermore, we would also like to thank Springer for publishing the proceedings in the prestigious series of *Lecture Notes in Computer Science.* We would, moreover, like to express our heartfelt appreciation to the keynote, plenary, panel, and invited speakers for their vision and discussions on the latest.

Finally, we would like to thank all the speakers, authors, and participants for their great contribution and support that made ICONIP 2013 a huge success.

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government.

November 2013

Minho Lee
Akira Hirose
Rhee Man Kil
Zeng-Guang Hou

Organization

Honorary Chair

Shun-ichi Amari	RIKEN, Japan
Soo-Young Lee	KAIST, Korea

General Chair

Minho Lee	Kyungpook National University, Korea

Program Chair

Akira Hirose	The University of Tokyo, Japan
Zeng-Guang Hou	The Chinese Academy of Sciences, China
Rhee Man Kil	Sungkyunkwan University, Korea

Organizing Chair

Hyeyoung Park	Kyungpook National University, Korea

Workshop Chair

Daijin Kim	POSTECH, Korea
Kyunghwan Kim	NT Research, Korea
Seong-Whan Lee	Korea University, Korea

Special Session Chair

Sung-Bae Cho	Yonsei University, Korea
Seiichi Ozawa	Kobe University, Japan
Liqing Zhang	Shanghai Jiao Tong University, China

Tutorial Chair

Seungjin Choi	POSTECH, Korea

Publication Chair

Yoonsuck Choe	Texas A&M University, USA
Hyung-Min Park	Sogang University, Korea
Seong-Bae Park	Kyungpook National University, Korea

Publicity Chair

Kazushi Ikeda	NAIST, Japan
Chi-Sing Leung	University of Hong Kong, Hong Kong
Shaoning Pang	Unitec Institute of Technology, New Zealand

Registration Chair

Min-Young Kim	Kyungpook National University, Korea

Financial Chair

Sang-Woo Ban	Dongguk University, Korea

Local Arrangement Chair

Doo-Hyun Choi	Kyungpook National University, Korea
Jong-Seok Lee	Yonsei University, Korea
Rammohan Mallipeddi	Kyungpook National University, Korea

Advisory Committee

Jonathan H. Chan, Thailand
Wlodzislaw Duch, Poland
Kunihiko Fukushima, Japan
Tom Gedeon, Australia
Aike Guo, China
Akira Iwata, Japan
Nik Kasabov, New Zealand
Irwin King, Hong Kong
Noboru Onishi, Japan
Ron Son, USA
Il Hong Suh, Korea
Shiro Usui, Japan
DeLiang Wang, USA
Lipo Wang, Singapore
Jun Wang, Hong Kong
Lei Xu, Hong Kong
Takeshi Yamakawa, Japan
Byoung-Tak Zhang, Korea
Li-Ming Zhang, China

Program Committee Members

Tani Jun
Soo-Young Lee
Sung-Bae Cho
Sungmoon jeong
Kyung-Joong Kim
C.K. Loo
Nung Kion Lee
Shan He
Dae-Shik Kim
Rubin Wang
Xin Yao
S. Ma
Honghai Liu
Joarder Kamruzzaman
Mallipeddi Rammohan
Zhirong Yang
Anto Satriyo Nugroho
Nikola Kasabov

Jonghwan Lee
Yaochu Jin
DaeEun Kim
Tingwen Huang
Fangxiang Wu
Dongbing Gu
Hongli Dong
Cesare Alippi
Kyung Hwan Kim
Lae-Jeong Park
Sang-Woong Lee
Sabri Arik
Chee-Peng Lim
Haibo He
Dat Tran
Kee-Eung Kim
Seungjin Choi
Robert (Bob) McKay
Xueyi (Frank) Wang
Jennie Si
Markus Koskela
Ivo Bukovský
Ryo Saegusa
El-Sayed El-Alfy
Hyeyoung Park
Bunthit Watanapa
Vinh Nguyen
Kalyana C. Veluvolu
Mufti Mahmud
Gil-Jin Jang
Hyung-Min Park
Jeounghoon Kim
Rhee Man Kil
Sang-Woo Ban
Masa Takatsuka
Chee Seng Chan
Pau-Choo Chung
Uvais Qidwai
Dong-Joo Kim
JongJin Lim
Sungoh Kwon
Long Cheng
Akitoshi Hanazawa
Andrzej Cichocki
Sheng Li
Oclay Kursun
Michel Verleysen
Peter Erdi
Qingsong Song
Bin Li
Huaguang Zhang
Derong Liu
Eric Matson
Mehdi Roopaei
Jacek Ma’ndziuk
Yang Shi
Zhiwu Lu
Xiaofeng Liao
Zhigang Zeng
Ding-Xuan Zhou
James Tin-Yau Kwok
Hsuan-Tien Lin
Osman Elgawi
Chao Zhang
Bo Shen
Nistor Grozavu
Younès Bennani
Jinde Cao
Li-Po Wang
Justin Dauwels
Andrew Leung
Bao-Liang Lu
Changyin Sun
Hong Yan
Abdesselam Bouzerdoum
Emili Balaguer Ballester
H. Tang
Roland Goecke
Jose Alfredo Ferreira Costa
Shri Rai
Kuntal Ghosh
TaeHyun Hwang
Alexander Rast
Yangming Li
Akira Hirose
Akira Iwata
Ko Sakai
Koichiro Yamauchi

Atsushi Shimada
Eiji Uchino
Hayao Shouno
Hayaru Shouno
Heizo Tokutaka
Hideki Asoh
Hiroaki Gomi
Hiroshi Dozono
Hiroshi Kage
Hiroshi Yamakawa
Hiroyuki Nakahara
Ikuko Nishikawa
Jinglu Hu
Jun Nishii
Katsumi Tateno
Katsunari Shibata
Kazuho Watanabe
Kazushi Ikeda
Kazuyuki Samejima
Keisuke Kameyama
Kenichi Tanaka
Kenichiro Miura
Kenji Doya
Kiyohisa Natsume
Masafumi Hagiwara
Michio Niwano
Mingcong Deng
Naoyuki Sato
Noboru Ohnishi
Seiichi Ozawa
Shin Ishii
Shinichiro Kano
Shunji Satoh
Shunshoku Kanae
Takashi Morie
Takashi Omori
Takeshi Aihara
Takio Kurita
Tao Ban
Tetsuya Yagi
Tohru Nitta
Tom Shibata
Toshiaki Omori
Toshihisa Tanaka
Yen-Wei Chen
Yoko Yamaguchi
Yoshikazu Washizawa

Table of Contents – Part III

Distance- and Direction-Dependent Synaptic Weight Distributions for Directional Spike Propagation in a Recurrent Network: Self-actuated Shutdown of Synaptic Plasticity

Toshikazu Samura[1], Yutaka Sakai[1], Hatsuo Hayashi[2], and Takeshi Aihara[1]

[1] Tamagawa University Brain Science Institute,
6-1-1 Tamagawa Gakuen, Machida, Tokyo 194-8610, Japan
{samura,sakai,aihara}@eng.tamagawa.ac.jp
[2] Kyushu Institute of Technology,
1-1 Sensui-cho, Tobata-ku, Kitakyushu, Fukuoka 804-8550, Japan
hayashi@brain.kyutech.ac.jp

Abstract. It has been suggested that directional spike propagation in a paradoxical way is organized in a recurrent network with anisotropic inhibition when excitatory connections to excitatory neurons (E–E) and inhibitory interneurons (E–I) are updated through spike-timing dependent plasticity. In this study, we show that both E–E and E–I connections have distance- and direction-dependent synaptic weight distributions in the recurrent network. E–E and E–I connections in the direction of spike propagation are more potentiated with increasing the distance between pre- and postsynaptic neurons. However, excitatory connections in the opposite direction of spike propagation are depressed regardless of the distance. In this network, the removal of the distance-dependency of E–I connections expands the width of directional spike propagation. On the other hand, the removal of the direction-dependency of E–I connections contracts spike propagation. These results show that the distance- and direction-dependent synaptic weight distributions contribute to directional spike propagation. The distance-dependent synaptic weight distribution, which suppresses activities in the lateral areas of the directional spike propagation, stops the progress of synaptic enhancement in those areas as if the synaptic plasticity is equipped with a self-actuated shutdown mechanism.

Keywords: recurrent network, spike propagation, STDP, distance- and direction-dependency, synaptic weight distribution.

1 Introduction

Yoshida and Hayashi have demonstrated that a hippocampal CA3 recurrent network organizes radial spike propagation from a stimulus site when recurrent excitatory connections between excitatory neurons (E–E) are updated by spike-timing dependent plasticity (STDP) [1]. The recurrent network consists of excitatory neurons and inhibitory interneurons. Both types of neurons are locally

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 1–8, 2013.

connected with other neurons. Although the radial spike propagation is organized in the CA3 network model, it has been observed that the directional propagation of neuronal activity occurs in the rat hippocampal CA1 [2]; it is quite possible that directional spike propagation in CA3 is projected to CA1. We have, therefore, demonstrated that a recurrent network causing anisotropic inhibition organizes directional spike propagation through STDP because axon arbors of O-LM cells (inhibitory interneurons) anisotropically spread in the hippocampus [3]. Furthermore, the modification of excitatory connections to inhibitory interneurons (E–I) alters the direction of spike propagation to the paradoxical direction where inhibitory interneurons have long inhibitory axons. These results imply that the organization of E–I connections is important for the directional spike propagation.

In this study, we show that the synaptic weights of excitatory connections (E–E and E–I connections) have different distributions depending on the direction of postsynaptic neurons with respect to the direction of spike propagation organized in the recurrent network. The synaptic weight distribution of excitatory connections to postsynaptic neurons in the direction of spike propagation depends on the distance between pre- and postsynaptic neurons. Long excitatory connections tend to acquire the maximum synaptic weight (distance dependency). On the other hand, synaptic weights of excitatory connections in the opposite direction of spike propagation tend to be weakened regardless of the distance between neurons; therefore synaptic weight distribution depends on the direction of postsynaptic neurons with respect to the direction of spike propagation (direction-dependency).

Furthermore, we removed either dependency of E–I connections by initializing synaptic weights of E–I connections in a recurrent network that organized paradoxical directional spike propagation. The removal of distance-dependency expands the width of directional spike propagation. On the other hand, the removal of direction-dependency contracts spike propagation. These results show that the direction-dependency of E–I connections allows a recurrent network to propagate spikes and the distance-dependency of E–I connections stops spike propagation in a direction perpendicular to the direction of spike propagation. The distance-dependent synaptic weight distribution, which suppresses activities in the lateral areas of the directional spike propagation, stops the progress of synaptic enhancement in those areas as if the synaptic plasticity is equipped with a self-actuated shutdown mechanism.

2 Methods

2.1 Recurrent Network

A recurrent network consists of excitatory neurons and inhibitory interneurons. Two types of neurons were a simple spiking model that was developed by Izhikevich [4]. The membrane potential of the ith neuron is calculated as follows:

$$v_i' = 0.04v_i^2 + 5.0v_i + 140.0 - u_i + I_i(t), \tag{1}$$

$$u_i' = a(bv_i - u_i), \tag{2}$$

where u_i is the membrane recovery variable, a is the rate of recovery, and b is the sensitivity of the recovery variable. $I_i(t)$ represents inputs from other neurons and external inputs to the ith neuron that are calculated as follows:

$$I_i(t) = \sum_{j}^{N_i} w_{ij} \sum_{k}^{N_j^{\text{fired}}} \delta(t - t_j^k - \tau_{ij}) + w^{\text{stim}} \sum_{l}^{N_i^{\text{stim}}} \delta(t - t_i^l), \tag{3}$$

where N_i is the number of presynaptic neurons of the ith neuron and w_{ij} represents synaptic weight between the ith and jth neurons. N_j^{fired} is the number of firing of the jth neuron. $\delta(\cdot)$ is the Dirac delta function and t_j^k is the kth firing timing of the $j$the neuron. τ_{ij} is a synaptic delay between the ith and jth neurons. w^{stim} is a synaptic weight for external inputs and N_i^{stim} is the number of external inputs to the ith neuron. t_i^l is the timing of external inputs. If v_i is larger than 30, the neuron fires and v_i and u_i are reset to c and $u_i + d$, respectively. Excitatory neurons were modeled as an intrinsic bursting neuron, so that we set parameters as follows: $a = 0.02, b = 0.2, c = -55, d = 5$. We modeled inhibitory interneurons as a fast spiking neuron, so that we set parameters as follows: $a = 0.1, b = 0.2, c = -65, d = 2$.

Figure 1 shows a part of network structure. $10,000$ excitatory neurons were placed on 100×100 lattice points. $1,250$ inhibitory interneurons were placed uniformly among excitatory neurons. Excitatory neurons were connected to surrounding 26 excitatory neurons and $1 - 6$ inhibitory interneurons randomly selected within each excitatory connectable region (9×7). On the other hand, inhibitory interneurons were connected to 48 excitatory neurons (I–E) randomly selected within each inhibitory connectable region (7×11). The connectable region of a neuron near the edge of the network was moved inside to keep the number of connectable neurons. Excitatory (E–E and E–I) connections had 2.0 ms synaptic delay, and E–E and E–I connections have respective initial synaptic weights of 4.0 and 3.0. Inhibitory (I–E) connections had 1.0 ms synaptic delay. The synaptic weights of I–E connections was -6.0, but the synaptic weight of interneurons near the edge was -18.0.

2.2 Synaptic Plasticity

Excitatory (E–E and E–I) connections were updated by STDP [5]. A spike of the ith neuron is paired with a arrival spike from the jth neuron in the nearest neighbor manner. In each spike pair, the modification rate of a synaptic weight from the jth neuron to the ith neuron is calculated as follows:

$$\Delta w_{ij} = \begin{cases} A_+ \ e^{-t_{ij}/\tau_{\text{STDP}}} & \text{if } \Delta\ t_{ij} > 0 \\ A_- \ e^{t_{ij}/\tau_{\text{STDP}}} & \text{if } \Delta\ t_{ij} \leq 0, \end{cases} \tag{4}$$

Δt_{ij} denotes an arrival timing of a spike from jth neuron relative to a spike timing of ith neuron. A_+, A_- are the maximal potentiation and depression rates respectively. τ_{STDP} is the time constant for STDP. We set these parameters as

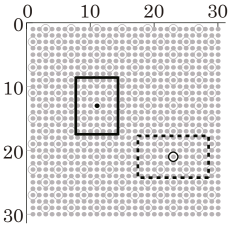


Fig. 1. A part of network structure. A gray dot is an excitatory neuron and a gray open circle is an inhibitory interneuron. The solid line box is an excitatory connectable region of a neuron (●). The dashed line box is an inhibitory connectable region of an interneuron (◯).

follows: $A_+ = 0.1$, $A_- = -0.12$, $\tau_{\mathrm{STDP}} = 20$ ms. Each synaptic weight was updated at each 1.0 ms. Synaptic weights are limited to the range of $1.0 \leq w \leq 6.0$.

3 Results

3.1 Directional Spike Propagation Organized in a Paradoxical Direction

In this simulation, we observed spike propagation organized in 30 trials. At the beginning of a trial, we initialized all connections and randomly chose 35 neurons within the central 15×15 region of the network (input region). The 35 neurons were fired by stimuli every 500 msec for $1,000$ sec.

Figure 2 shows firing probability during simulation in 30 trials. First, neurons in the central region were activated by stimulation (Fig. 2 (a)). The activities, then, expanded to surrounding region especially in the horizontal direction (Fig. 2 (b)). Finally, the expansion of the activities became steady because the difference between Fig. 2 (b) and (c) is small. The activities evoked by stimuli always propagated to the right and left sides from the input region. Thus, this network organized spike propagation paradoxically in the horizontal direction where inhibitory interneurons have long inhibitory axons.

3.2 Distance- and Direction-Dependent Synaptic Weight Distributions

Radial spike propagation was organized in the network by decreasing the number of inhibitory connections (48 → 28). Under the condition, we obtained synaptic weight distributions from neurons in the area that does not include the edge of the network and the input region. Here, we divided excitatory connections into two groups depending on the direction of postsynaptic neurons. One group consists of excitatory connections to postsynaptic neurons in the direction of spike propagation. The other group consists of excitatory connections to postsynaptic

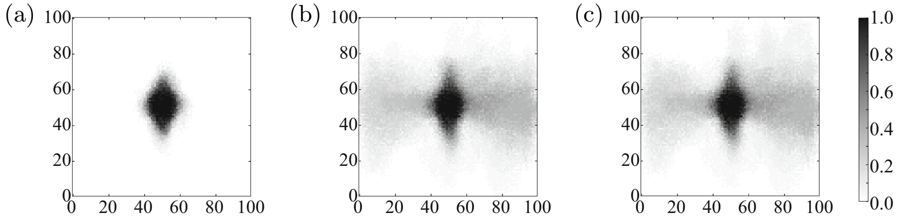


Fig. 2. The organization of horizontal spike propagation. Firing probability of each excitatory neuron in 30 trials are calculated at 50 (a), 750 (b), and 1000 (c) sec.

neurons in the opposite direction of spike propagation. The direction of spike propagation is different from neuron to neuron depending on its location and organized spike propagation in the network.

Synaptic weights of excitatory connections (E–E and E–I connections) have distributions depending on the direction of postsynaptic neurons (Fig. 3). Synaptic weights of excitatory connections depend markedly on the distance between pre- and postsynaptic neurons in the direction of spike propagation (Fig. 3 (a), (b)). Longer excitatory connections tend to be potentiated. On the other hand, excitatory connection in the opposite direction of spike propagation have almost no distance dependency (Fig. 3 (c), (d)). Many excitatory connections were depressed regardless of the distance between neurons in this direction; therefore, these synaptic weights depend only on the direction of postsynaptic neurons with respect to the direction of spike propagation.

3.3 Removal of Direction- and Distance-Dependent Synaptic Weight Distributions

It is inferred from the previous study [3] that organizing E–I connections is important for the directional spike propagation because it changes the direction of spike propagation. Therefore, we investigated effects of the removal of the distance- and direction-dependencies of E–I connections on the spike propagation.

We collected 10 trials that finally organized steady horizontal spike propagation reaching both sides of the network. Figure 4 (a) shows the firing probability of excitatory neurons during the steady horizontal spike propagation. Figure 4 (b) shows the firing probability of developing spike propagation. The dark region where neurons highly fire is wider than that of the steady propagation. These results indicate that the network organizes wide spike propagation before the propagation reaches a steady state.

We removed each dependency of E–I connections by initializing the synaptic weights of E–I connections of neurons that fired for the developing period. After that, we applied an input that caused a steady spike propagation in each trial and we confirmed the changes of the spike propagation caused by the input.

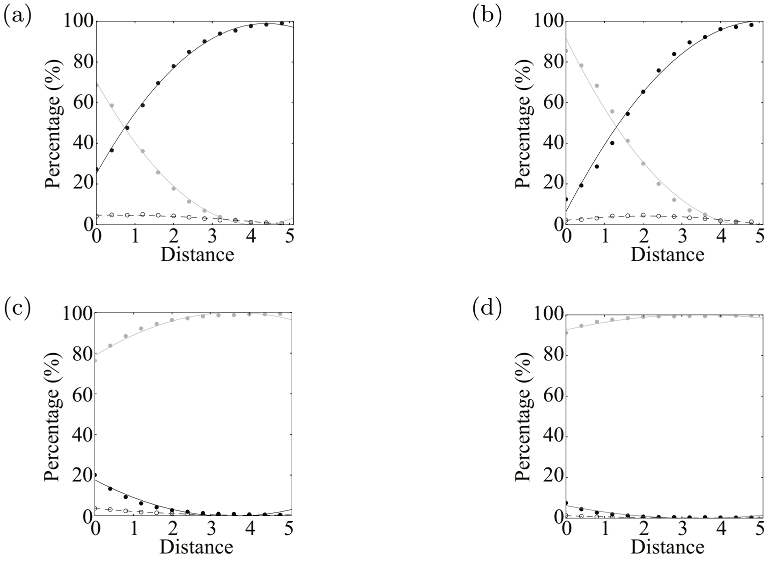


Fig. 3. Distance- and direction-dependent synaptic weight distributions. The ordinates are percentages of the maximum (black circle), the minimum (gray circle), and the intermediate (open circle) synaptic weights in each distance bin (the bin width was 0.4). Distance is the length of orthogonal projection of a connection between pre- and postsynaptic neurons onto the direction of spike propagation. E–E (a) and E–I (b) connections in the spike propagation direction have distance-dependent weight distribution. E–E (c) and E–I (d) connections in the opposite direction of spike propagation have almost no distance-dependency.

Figure 4 (c) shows the firing probability in the recurrent network where the direction-dependency is removed by initializing E–I connections in the opposite direction of spike propagation. Spike propagation organized in the network shrank compared to the steady spike propagation. The direction-dependency of E–I connections in the opposite direction of spike propagation allows the network to propagate spikes. Figure 4 (d) shows the firing probability in the recurrent network where the distance-dependency is removed by initializing E–I connections in the direction of spike propagation. In this case, the input activated not only the same neurons in the steady propagation but also other neurons that were not fired in the steady propagation. Spike propagation became wider than the steady propagation. The dark region of Fig. 4 (d) is similar to that of Fig. 4 (b). Therefore, acquiring distance-dependency of E–I connections makes spike propagation narrow as shown in Fig 4 (b) → (a). This means that the distance-dependency of E–I connections in the direction of spike propagation prevent the network from propagating spikes in a direction perpendicular to the direction of spike propagation.

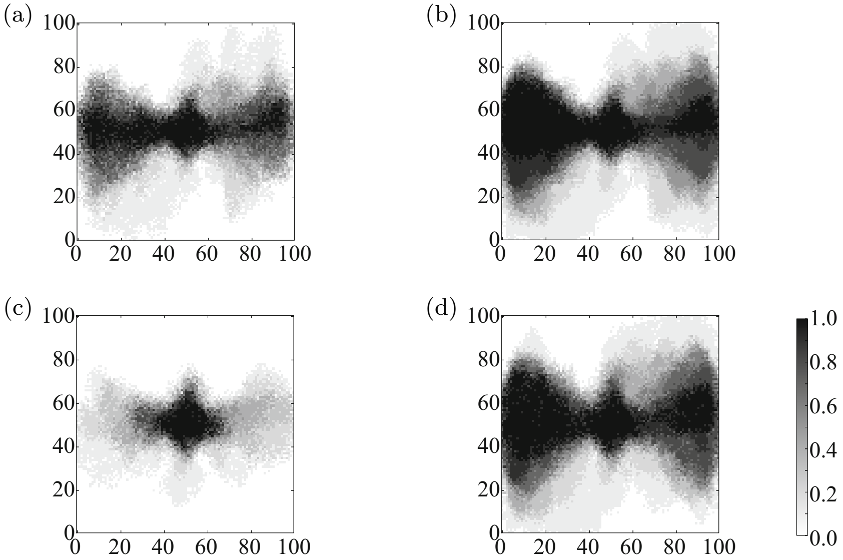


Fig. 4. Firing probability of excitatory neurons in 10 trials calculated from steady spike propagation (a), developing spike propagation (b), and spike propagation after the removal of direction dependency (c) or distance dependency (d)

4 Conclusion and Discussion

Directional spike propagation in a paradoxical way is organized in a recurrent network with anisotropic inhibition when E–E and E–I connections are updated through STDP.

We found that E–E and E–I connections have different distributions depending on the direction of postsynaptic neurons with respect to the direction of spike propagation in the recurrent network organizing spike propagation. Excitatory connections in the direction of spike propagation have a distance-dependent synaptic weight distribution. On the other hand, excitatory connections in the opposite direction of spike propagation have no distance-dependency. We found that the removal of the direction- and distance-dependencies of E–I connections shrink and widen spike propagation, respectively. The direction-dependency allows a recurrent network to cause spike propagation and the distance-dependency restricts the range of spike propagation in a direction perpendicular to the direction of spike propagation. Consequently, distance- and direction-dependent synaptic weight distributions contribute to cause directional spike propagation in a recurrent network.

The distance-dependency of E–I connections prevents a network from propagating spikes in a direction perpendicular to the direction of spike propagation. Synaptic plasticity progresses no further in this direction because no neuronal activities occur in this direction. Thus, as if synaptic plasticity had mechanisms of self-actuated shutdown, synaptic plasticity switches itself on and off according to the progress of synaptic change. Indeed, excitatory connections to inhibitory interneurons are updated through synaptic plasticity in the brain (reviewed in [6]). The automatic shutdown mechanisms of synaptic plasticity might be equipped in the brain; therefore, we should also focus on the reorganization of E–I connections in ongoing activities to understand the function in the brain activity.

Acknowledgements. This study was supported by MEXT -Supported Program for the Strategic Research Foundation at Private Universities, 2009–2013.

References

1. Yoshida, M., Hayashi, H.: Regulation of Spontaneous Rhythmic Activity and Organization of Pacemakers as Memory Traces by Spike-Timing-Dependent Synaptic Plasticity in a Hippocampal Model. Phys. Rev. E 69, 011910:1–011910:15 (2004)
2. Lubenov, E.V., Siapas, A.G.: Hippocampal Theta Oscillations Are Travelling Waves. Nature 459, 534–539 (2009)
3. Samura, T., Hayashi, H.: Directional spike propagation in a recurrent network: Dynamical firewall as anisotropic recurrent inhibition. Neural Netw. 33, 236–246 (2012)
4. Izhikevich, E.M.: Simple Model of Spiking Neurons. IEEE Trans. Neural Netw. 14, 1569–1572 (2003)
5. Bi, G.Q., Poo, M.M.: Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J. Neurosci. 18, 10464–10472 (1998)
6. Kullmann, D.M., Lamsa, K.P.: LTP and LTD in cortical GABAergic interneurons: Emerging rules and roles. Neuropharmacology 60, 712–719 (2011)

Modulated Neuronal Activity and Connectivity of Smoking Resist Using Real-Time fMRI Neurofeedback

Dong-Youl Kim and Jong-Hwan Lee*

Department of Brain and Cognitive Engineering, Korea University
Seoul 136-713, Korea
{kdy1984,jonghwan_lee}@korea.ac.kr

Abstract. Recent functional magnetic resonance imaging (fMRI) technique with real-time (rt) feedback has widely been adopted to regulate one's own neuronal activity within regions-of-interest (ROIs). Despite the fact that the functional connectivity (FC) between ROIs has also been modulated via rt-fMRI neurofeedback (NF), however there is no study to explicitly provide the FC patterns in addition to neuronal activity levels during rt-fMRI NF trials. In this study, we adopted both neuronal activities within an ROI and FC patterns between ROIs to investigate a potential utility of the FC information. Fourteen heavy smokers could voluntarily control their brain activity based on the neurofeedback of both neuronal activation within an ROI related to smoking resist and FC patterns between ROIs. Our proposed rt-fMRI method appears to modulate not only the neuronal activity but also the neuronal connectivity levels.

Keywords: Functional magnetic resonance imaging, smoking resist, real-time fMRI neurofeedback, orbitofrontal cortex, anterior cingulate cortex, posterior cingulate cortex, precuneus, functional connectivity.

1 Introduction

Functional magnetic resonance imaging (fMRI) modality has been widely used to explore various functions of human brain in a noninvasive manner [1]. Recently, multiple fMRI research groups have investigated the feasibility of regulating the brain activity using real-time (rt) neurofeedback (NF) [2, 3]. Using rt-fMRI NF technique, the potential therapeutic benefit has been explored for a number of disorders including the chronic pain [4], Parkinson's disease [5], schizophrenia [6], and major depression [7]. Moreover, a feasibility of rt-fMRI NF for treatment of addictive disorders including nicotine dependence has been demonstrated [8].

In these previous studies, it has also reported that the functional connectivity (FC) of brain regions could be modulated via rt-fMRI NF adopting feedback information based on neuronal activity of a region-of-interest (ROI). However, there is no study to explicitly provide the FC as well as neuronal activity levels as feedback information in the rt-fMRI NF.

[*] Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 9–16, 2013.

We hypothesized that the neuronal activity and connectivity patterns related to smoking resist would efficiently be modulated through rt-fMRI NF based on the feedback signal consisted of both neuronal activity and FC compared to the conventional approach employing the neuronal activity only.

2 Materials and Methods

2.1 Participants and Imaging Parameters

Fourteen right-handed male heavy smokers without any neurological and neuropsychiatric disease were recruited. Inclusion criteria were: the Fagerström Test of Nicotine Dependence scores (> 4, 4.86±0.95; [9]); years of smoking (> 5 years, 7.36±1.91); and number of cigarettes per day (> 10, 16.50±2.93). The expired-air carbon oxide (CO) levels (piCO smokerlyzer; Bedfont Scientific, Ltd., Rochester, UK) were measured in normal condition (21.21±4.14) and before each of two fMRI sessions with at least 6 hours cessation (session1: 8.43±3.16, session2: 9.57±3.41). The blood-oxygenation-level-dependent (BOLD) fMRI scans were acquired from all participants using a Siemens Tim Trio 3-T scanner (Erlangen, Germany). Functional data were obtained using a standard T_2^*-weighted gradient-echo echo-planar-imaging (EPI) pulse sequence from the whole brain (TR/TE = 1000/24 ms; FoV = 24×24 cm^2; matrix size = 64×64; voxel size = 3.75×3.75×7.0 mm^3; FA = 90°; interleaved 20 axial slices with no gap).

2.2 Real-Time fMRI (rt-fMRI) NF Method

Figure 1 shows the overall rt-fMRI NF method including steps to acquire raw fMRI signals from workstation of MR control, to analyze the acquired data in real-time, and to present the NF to the participants through visual goggles. The reconstructed EPI volumes in MRI workstation were transferred to a laptop computer via TCP/IP file transfer protocol. The received fMRI data were preprocessed including the head motion corrections and temporal smoothing across the 3TR time points before extracting the feedback signal in the rt-fMRI NF trials.

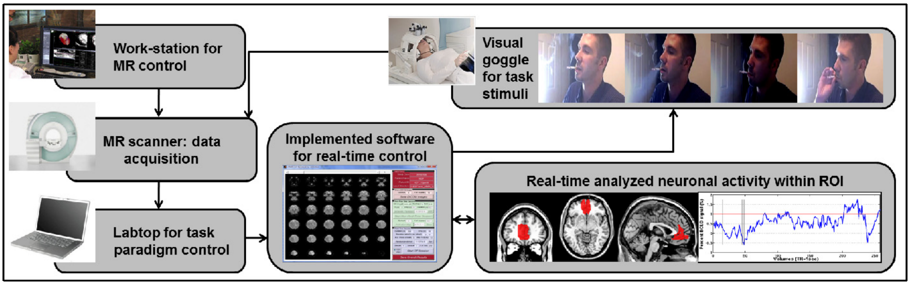

Work-station for MR control
MR scanner: data acquisition
Labtop for task paradigm control
Implemented software for real-time control
Visual goggle for task stimuli
Real-time analyzed neuronal activity within ROI
Volumes (TR=1sec)

Fig. 1. An illustration of real-time fMRI NF setup

2.3 Experimental Procedure

In each of two scanning sessions, four non-real-time runs were performed as designed with block-based paradigm using smoking-related images for all participants (two runs before rt-fMRI runs and two runs after rt-fMRI runs). For each non-real-time run (5-min 20-sec), five blocks for each of smoking and neutral scenes were counter-balanced and interleaved with ten fixation blocks (5-images per block; 3-sec per image). All participants underwent each of non-real-time runs with the instruction to 'allow oneself to crave' or to 'resist the urge to smoke'. Two conditions of craving (C) and resistance (R) to smoke are counter-balanced before and after rt-fMRI runs (*i.e.*, CR…rt-fMRI runs…RC).

Two smoking-resist related ROIs were defined as the (1) bilateral medial orbitofrontal cortex and anterior cingulate cortex (ROI1; [10, 11] and (2) bilateral precuneus and posterior cingulate cortex (ROI2; [10, 11]). At the start of each rt-fMRI scan, first acquired EPI volume was normalized to Montreal Neurological Institute (MNI) template and then two ROIs were warped to normalized EPI. Seven participants were included in the first group (FB1) and they were provided the NF signal solely based on neuronal activity level within an ROI1. The remaining seven participants were included in the second group (FB2) and they were provided a NF signal consisted of both the neuronal activity in an ROI1 and FC between the ROI1 and ROI2. Participants in the FB1 and FB2 were randomly assigned from all 14 volunteers and were matched in their demographic information. Six rt-fMRI scan runs were acquired after at least 6 hours of smoking cessation during two separate visits in one week apart. Each run was consisted of 15-sec calibrations, 30-sec cross fixation, 3-sec ready instruction, 180-sec video stimuli with smoking scenes, 10-sec craving rate question, and 20-sec cross fixation (258-sec for each run). Twelve video stimuli consisted of 3 minutes of smoking scenes were pseudo-randomly played to each participant and delivered to the participants via MR-compatible visual goggles (NordicNeuroLab Inc., www.nordicneurolab.com).

In a real-time NF run, the NF signal was represented as contrast changes of the smoking-related video clip display. In detail, participants were instructed to attempt to black out the screen, i.e. the screen becomes darker when the feedback signal is increased and vice versa. For participants in the FB1 group, the percentage BOLD (PB) intensity of the ROI1 was calculated and then used as a feedback signal. The averaged BOLD signal across the voxels within the ROI1 was band-pass filtered with range from 0.008 to 0.1 Hz and was detrended. To estimate the level of PB of the ROI1, the period of cross fixation with delay of 6-sec from 21-sec to 45-sec was determined as the baseline of BOLD signal. The level of PB was continuously updated as an averaged BOLD signal across recent 3 volumes.

For the participants in the FB2 group, both the neuronal activity in the form of the PB level of the ROI1 and the FC level between the two ROIs was used as feedback signal. The FC between the average BODL signals of the ROI1 and ROI2 was calculated using Pearson's correlation coefficients.

2.4 Preprocessing and Data Analysis

The acquired BOLD fMRI data of each rt-fMRI run were preprocessed using realignment to the head motion correction and spatial smoothing using an 8 mm isotropic FWHM Gaussian kernel in SPM8 toolbox (www.fil.ion.ucl.ac.uk/spm). The first 45 volumes during the calibration and cross-fixation were excluded from raw BOLD fMRI data and the preprocessed fMRI data were band-pass filtered with the range from 0.01 to 0.08 Hz and then were detrended.

In the individual level analysis, the level of PB was calculated voxel-wise using the preprocessed fMRI data and an average of lower 10 percentile values across the video stimulus presentation was adopted as baseline level. The criterion which is to maintain at least three seconds (*i.e.*, 3 TR) was applied to select those values. This baseline level was adopted in our study due to an inability to define the resting period in the real-time feedback trial during 180-sec (i.e. the lower 10 percentiles values were considered to be baseline levels) [12]. The top 10 percentile PB intensity values were used to calculate their increased BOLD intensity based on the NF trials. The averaged PB patterns were normalized by z-scoring across voxels within an ROI. To investigate the feasibility of modulation effect, the mean value within a cluster (z > 1.96) was estimated for each run of each session.

To estimate the FC level, the averaged time-series within each of ROIs were used as the reference time-series in each of the two ROIs. The Pearson's correlation coefficients were calculated between two reference time-series for each run of each subject.

3 Results

3.1 Self-modulation Effects of Neuronal Activity

Figure 2 represents the mean and standard error of active patterns from the averaged PB within each of the two ROIs across subjects. A paired *t*-test was conducted for investigating the difference of neuronal activity between two sessions. On the domain

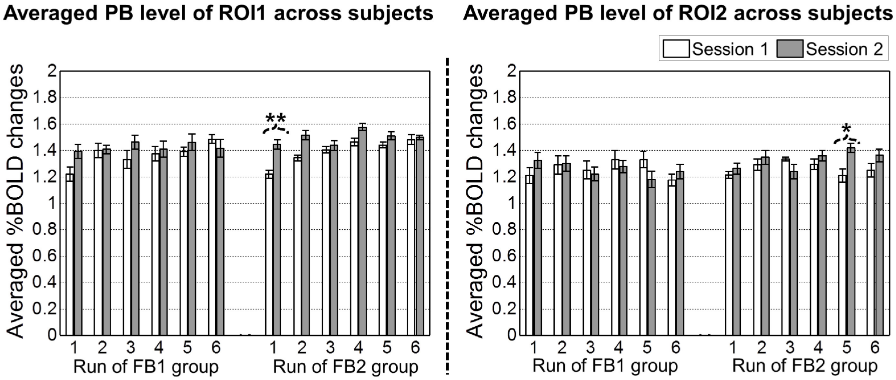


Fig. 2. Averaged level of percent BOLD signals of each of two ROIs across all subjects

of each rt-fMRI run for ROI1, significant difference was estimated only in first run of FB2 (p=3.66×10^{-5}) and moderate differences were found in first run of FB1 (p=0.06) and second run of FB2 (p=0.08; p of other cases > 0.1). For the neuronal activity in ROI2, only the fifth run in FB2 group showed significant difference between two sessions (p=0.05).

Group-level analysis was performed using a paired t-test between FB1 and FB2 groups of all six runs for each session and between sessions of all seven subjects for each group as shown in Figure 3 (mean±standard error of ROI1: 1.37±0.05% for FB1 in the first session, 1.42±0.05% for FB1 in the second session, 1.39±0.01% for FB2 in the first session, and 1.49±0.02% for FB2 in the second session). Between two groups for ROI1, no significant difference was reported for each of two sessions. On the other hand, the significant session difference was found in FB2 group (p=0.01), but not in FB1 group (p=0.18). For ROI2, the neuronal activity from averaged PB across runs (mean±standard error: 1.25±0.01 and 1.26±0.01% for FB1 in the first and second session, respectively, 1.26±0.01 and 1.33±0.01% for FB2 in the first and second session, respectively) was significantly different between sessions in FB2 group (p=0.02) but not in FB1 group (p=0.89).

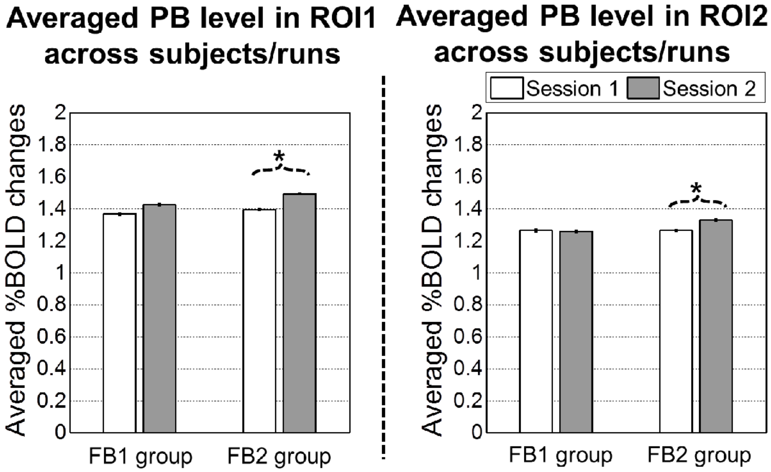


Fig. 3. Averaged level of percent BOLD signals of each of two ROIs across subjects/runs

3.2 Effect of Functional Connectivity between Two ROIs

Figure 4 illustrates the individual level of FC between averaged time-series of two ROIs (mean±standard deviation of correlation coefficients: 0.58±0.16 and 0.60±0.22 for FB1 in the first and second sessions, respectively, 0.65±0.18 and 0.66±0.18 for FB2 in the first and second sessions, respectively). From a paired t-test result, there was no difference between two sessions for each group (all p > 0.5). It was interesting that the significant group difference (FB2 > FB1 group) was found in the first session (p=0.04) but not in the second session (p=0.19).

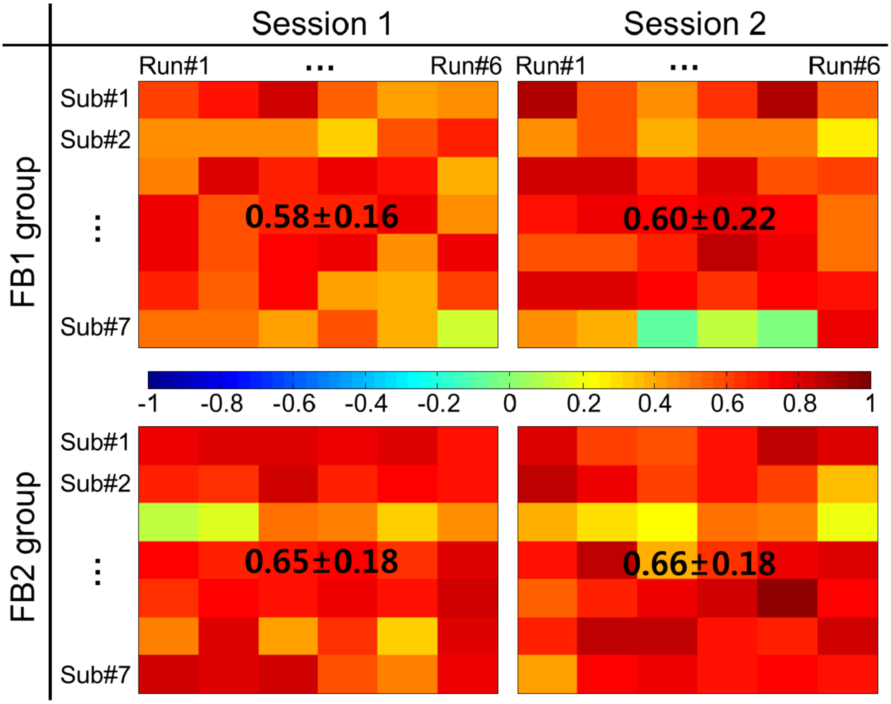


Fig. 4. Functional connectivity between two ROIs for all subjects/runs

4 Discussion

The results found in this study suggest that the self-modulation can be improved through repeated runs in the first visit but not the second visit regarding the neuronal activity. The degree of modulation from neuronal activity in the first (both FB1/FB2) and second runs (only FB2) are different between two sessions but the level of PB are comparable between sessions in latter runs. These results may indicate that the self-modulation is fully learned in the first visit, so the self-modulation effect does not increase further in the second visit but maintained. Another intriguing fact is that the meaningful difference of neuronal activity between sessions is not reported for FB1 but for FB2 group. This result may suggest that the degree of self-modulation is affected by the type of NF using both neuronal activity and FC. Even though the degree of FC was not different between two sessions for each group, the relatively stronger connection was estimated in FB2 compared to FB1 group. Hence, if the feedback signal is generated considering both neuronal activity and FC between an ROI and the ROI-connected brain regions, the effect of modulation of brain function would be increased through repeated real-time fMRI runs.

One of the most important issues in treating nicotine dependence is that the ability to modulate smoking desire in MRI scanner can be applied to the natural environments [8]. With the limitation in mind, future works are warranted to investigate the modulated neuronal patterns within ROIs using acquired non-rtfMRI scans before/after rt-fMRI scans and the potential facilitation of self-modulation from rt-fMRI neurofeedback scans. Compared to previous rt-fMRI studies [3, 4, 8], we showed the increased performance using NF signals as both neuronal activity and FC patterns. The proposed of rt-fMRI NF method constituting feedback signal using the FC levels in addition to the neuronal activity appears to benefit to efficiently regulate neuronal networks compared to the method based on the feedback signal using the neuronal activity level based on our preliminary analysis.

It is important to note that we observed that the BOLD signals of the voxels within the ROI1 and ROI2 were shown substantial fluctuations even between the voxels in proximity. Based on our evaluation, it seems that there are non-neuronal artifacts including the head motions and physiological artifacts dominant in the white matter and cerebro-spinal fluids were confounded in the BODL signal used in this report. Thus, it would be an interesting further study, how these non-neuronal components could alter the analytic results from the data measured in our study by removing these non-neuronal components via least-squares based denoising method. The conclusive evidence on the efficacy of our proposed rt-fMRI NF scheme is pending until the systematic analysis is conducted using the artifact reduced BOLD signals. The important other is that the observed BOLD signals have the different resting and active periods depending on the voxels even the voxels are located in same ROI. Therefore, it should be managed in future study that the resting and active periods are not localized for all voxels within ROI.

5 Conclusion

In this study, we presented the feasibility of alteration of neuronal networks using real-time fMRI NF scheme using both the neuronal activity and functional connectivity levels as feedback signal. Further investigation is warranted to justify the efficacy of the proposed scheme via systematic comparisons with the conventional rt-fMRI NF scheme based solely on the neuronal activity level.

Acknowledgments. This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (MEST) (2012-0002342) and in part by Samsung Electronics, Inc.

References

1. Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A.: Neurophysiological Investigation of the Basis of the fMRI Signal. Nature 412, 150–157 (2001)
2. LaConte, S.M.: Decoding fMRI brain states in real-time. Neuroimage 56, 753–765 (2011)
3. Soldati, N., Calhoun, V.D., Bruzzone, L., Jovicich, J.: The Use of a Priori Information in ICA-based Techniques for Real-Time fMRI: an Evaluation of Static/Dynamic and Spatial/Temporal Characteristics. Front. Hum. Neurosci. 7, 64 (2013)

4. de Charms, R.C., Maeda, F., Glover, G.H., Ludlow, D., Pauly, J.M., Soneji, D., Gabrieli, J.D., Mackey, S.C.: Control over Brain Activation and Pain Learned by Using Real-Time Functional MRI. Proc. Natl. Acad. Sci. U. S. A. 102, 18626–18631 (2005)
5. Subramanian, L., Hindle, J.V., Johnston, S., Roberts, M.V., Husain, M., Goebel, R., Linden, D.: Real-Time Functional Magnetic Resonance Imaging Neurofeedback for Treatment of Parkinson's Disease. J. Neurosci. 31, 16309–16317 (2011)
6. Ruiz, S., Lee, S., Soekadar, S.R., Caria, A., Veit, R., Kircher, T., Birbaumer, N., Sitaram, R.: Acquired Self-Control of Insula Cortex Modulates Emotion Recognition and Brain Network Connectivity in Schizophrenia. Hum. Brain. Mapp. 34, 200–212 (2013)
7. Linden, D.E., Habes, I., Johnston, S.J., Linden, S., Tatineni, R., Subramanian, L., Sorger, B., Healy, D., Goebel, R.: Real-Time Self-Regulation of Emotion Networks in Patients with Depression. PLoS One 7, e38115 (2012)
8. Hanlon, C.A., Hartwell, K.J., Canterberry, M., Li, X., Owens, M., Lematty, T., Prisciandaro, J.J., Borckardt, J., Brady, K.T., George, M.S.: Reduction of Cue-Induced Craving through Realtime Neurofeedback in Nicotine Users: The Role of Region of Interest Selection and Multiple Visits. Psychiatry. Res. 213, 79–81 (2013)
9. Heatherton, T.F., Kozlowski, L.T., Frecker, R.C., Fagerström, K.O.: The Fagerström Test for Nicotine Dependence: a Revision of the Fagerström Tolerance Questionnaire. Br. J. Addict. 86, 1119–1127 (1991)
10. Hartwell, K.J., Johnson, K.A., Li, X., Myrick, H., LeMatty, T., George, M.S., Brady, K.T.: Neural Correlates of Craving and Resisting Craving for Tobacco in Nicotine Dependent Smokers. Addict. Biol. 16, 654–666 (2011)
11. Lee, J.H., Kim, D.Y., Kim, J.: Mesocorticolimbic Hyperactivity of Deprived Smokers and Brain Imaging. Neuroreport 23, 1039–1043 (2012)
12. Van De Ville, D., Jhooti, P., Haas, T., Kopel, R., Lovblad, K.O., Scheffler, K., Haller, S.: Recovery of the Default Mode Network after Demanding Neurofeedback Training Occurs in Spatio-Temporally Segregated Subnetworks. Neuroimage 63, 1775–1781 (2012)

Parameterized Digital Hardware Design of Pulse-Coupled Phase Oscillator Model toward Spike-Based Computing

Yasuhiro Suedomi, Hakaru Tamukoh, Michio Tanaka, Kenji Matsuzaka, and Takashi Morie

Graduate School of Life Science and Systems Engineering,
Kyushu Institute of Technology, Kitakyushu, 808-0196 Japan
suedomi-yasuhiro@edu.brain.kyutech.ac.jp, morie@brain.kyutech.ac.jp

Abstract. This paper proposes a parameterized digital circuit design approach for pulse-coupled phase oscillators. Our approach aims to construct a reconfigurable hardware platform that emulates a large-scaled pulse-coupled network with complicated interconnection toward spike-based computing. The network, which is described by the parameterized Verilog-HDL, can change the calculation accuracy, the coupling function shape of oscillators, the network size and interaction between oscillators by parameters. Experimental results show that a prototype designed by the proposed approach emulates well in-/anti-phase and different (out-of-phase) synchronization.

Keywords: pulse-coupled phase oscillator, synchronization, FPGA, parameterized hardware design.

1 Introduction

As the first step to construct intelligent information processing systems mimicking the brain architecture, which we call brain-inspired systems, we have proposed analog computation approaches with continuous-time nonlinear dynamics and time-domain computation using pulse-modulation signals or asynchronous spike pulses [4,6,7]. Such analog computation can be modeled by using pulse-coupled phase oscillator systems [9].

In a pulse-coupled phase oscillator system, mutual interactions are represented by phase sensitivity functions, and different functions lead to different synchronization phenomena. Coupling between oscillators are determined by the timing of pulses output from each oscillator. Based on this principle, because nonlinear information processing is performed by a single state transition at spike timing, high-performance and low-power intelligent information processing VLSI systems can be constructed.

In order to implement spike-based computation and to show its effectiveness, we have designed a CMOS circuit using the pulse-coupled phase oscillator, and applied it to a coupled Markov Random Field (MRF) model. The coupled MRF

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 17–24, 2013.

models provide practical algorithms for detecting discontinuities in motion, intensity, color, and depth in image scenes [2,3,5]. It can be applied to region-based image segmentation and implemented with pulse-coupled phase oscillators. The proposed circuit introduced a simplified region-based coupled MRF model for efficient hardware implementation [6]. By circuit and numerical simulations, we demonstrated that the proposed simplified MRF model not only retains the advantages of our old model [4], but also improves the processing speed and the region segmentation performance. From the above previous works, we confirmed the effectiveness of spike-based computing implemented by the merged analog/digital circuits. It is a promising approach for direct modeling of spike-based computation in the brain.

On the other hand, simulating very large-scaled pulse-coupled networks with complicated interconnection is desired to analyze realistic behavior of spike-based computation in the brain. Although the merged analog/digital circuit by means of ASIC implementation is appropriate to realizing neuromorphic VLSI, it has a drawback from the reconfigurability point of view to change the network size and network interconnection. Therefore, a reconfigurable platform to simulate various sizes and interconnections of pulse-coupled networks is required.

In this work, we propose a parameterized digital circuit design approach of pulse-coupled phase oscillators. The proposed design can change the calculation accuracy, the coupling function shape, the network size and interaction between neurons by parameters. It is described by the parameterized Verilog-HDL (Hardware Description Language) and run on a Field Programmable Gate Array (FPGA). Experimental results show that the proposed circuits well emulate pulse-coupled oscillator systems, and generates in-/ anti-phase and different (out-of-phase) synchronization.

2 Pulse-Coupled Phase Oscillator Model with Three-Valued Coupling Functions

The concept of coupled phase oscillators was firstly proposed by Winfree [9] and its dynamics is expressed as follows:

$$\frac{d\phi_i}{dt} = \omega_i + Z(\phi_i)Spk(t), \tag{1}$$

where ϕ_i is the i-th phase variable with 2π periodicity, ω_i the i-th natural angular frequency, $Z(\phi_i)$ the phase sensitivity function, which gives the response of the i-th oscillator. Inputs from other oscillators, $Spk(t)$, are assumed here the pulse inputs as follows:

$$Spk(t) = \frac{K_0}{N}\sum_{j=1}^{N}\sum_{n=1}^{\infty}\delta(t - t_{jn}), \tag{2}$$

where K_0 is the coupling strength, N the number of oscillators, and t_{jn} a firing time. Function δ is mathematically Dirac's delta function, and represents input spike timing without a pulse width. However, in real hardware, a spike pulse has

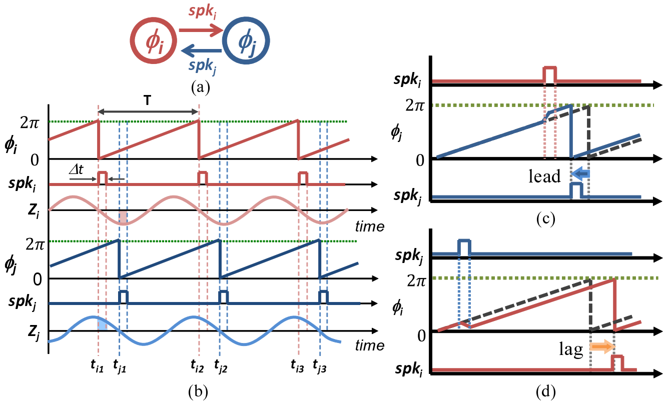


Fig. 1. Dynamics of pulse-coupled phase oscillators when $Z(\phi_i) = -\sin(\phi_i)$

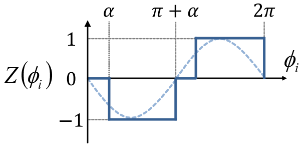


Fig. 2. Example of the shape of function $Z(\phi_i)$

a definite width Δt, during which ϕ_i is updated according to the value of $Z(\phi_i)$. As a specific case, if it is assumed that $Z(\phi_i) = -\sin(\phi_i)$, the dynamics of the pulse-coupled oscillators are schematically illustrated in Fig. 1.

In order to simplify the function shape of $Z(\phi_i)$ for hardware implementation, we have used three-value functions as $Z(\phi_i)$; a typical function shape is shown in Fig. 2, which is expanded in the time domain and has values: $-1, 0, 1$ [7]. In the update operation, "-1" and "1" correspond to decrease and increase of ϕ_i, respectively, and "0" means no updating. Increasing ϕ_i results in leading of next spike firing timing and vice versa.

We have also introduced parameter α as a span during which $Z(\phi_i) = 0$, as shown in Fig. 2 [7]. By setting $\alpha > \Delta t$, we can prevent connected oscillators from over-updating when they approach to a synchronization state. Compared with sinusoidal functions, the three-value functions lead to faster convergence and simpler circuit implementation.

3 Parameterized Digital Hardware Design of Pulse-Coupled Phase Oscillators

A circuit architecture of a pulse-coupled phase oscillator is shown in Fig. 3(a). It consists of an oscillator circuit, a function generator circuit and an update circuit.

The oscillator circuit consists of an n-bits counter (CNT), a spike generator (SPK_GEN) and combinational circuits, as shown in Fig. 3(b). The n-bits counter represents phase variable ϕ_i and counts the clock inputs for realizing ω_i in Eq. (1). The spike generator outputs a spike pulse $Spike_i$ to other oscillators when the n-bits counter reaches to the maximum value. The oscillator circuit also outputs $cMSB$, $cMid0$ and $cMid1$ which determine the shape of function $Z(\phi_i)$ and are used in the function generator circuit.

The function generator circuit combines signals $cMSB$, $cMid0$ and $cMid1$ into Z_p and Z_n, as shown in Fig. 3(c). The update circuit receives Z_p, Z_n and spike pulses $Spike_j$ received from other oscillators and outputs signal *update*, as shown in Fig. 3(d). It updates the value of the counter in the oscillator circuit for realizing lead and lag operations.

Figure 4 shows waveforms related to function $Z(\phi_i)$ generated by the function generator. If sign signal $sign = 0$, then signals Z_p and Z_n are output as the simplified function of $-\sin(x)$ shown in Fig. 2. On the other hand, if $sign = 1$, they are output as $+\sin(x)$.

In the proposed design, counter size n and span parameter α can be changed to control the calculation accuracy and the span during when $Z(\phi_i) = 0$, respectively. The following is a description on parameterization by Verilog-HDL that was used in the experiments.

```
parameter CNT_SIZE = 6;
parameter alpha = 1;
assign cMSB = cnt[CNT_SIZE-1];
assign cMid0 = |cnt[CNT_SIZE-2:alpha];
assign cMid1 = &cnt[CNT_SIZE-2:alpha];
```

4 Experimental Results

4.1 Logic Simulation

We simulated the pulse-coupled phase oscillator circuit system, and observed synchronization phenomena with a clock frequency of 200 MHz, where the simplified three-valued function shown in Fig. 4 was used.

Results of synchronization phenomena are shown in Fig. 5. When $sign = 0$ and the initial phase difference was 0.96π, two oscillators synchronized with an in-phase mode as shown in Fig. 5(a). On the other hand, when $sign = 1$ and the initial phase difference was 0.03π, two oscillators synchronized with an anti-phase mode as shown in Fig. 5(b).

In addition, out-of-order synchronizations were observed if we set $\alpha = 3$. When $sign = 0$ and the initial phase difference was 0.78π, two oscillators came close to each other but did not synchronize with an in-phase mode as shown in Fig. 5(c). Similarly, when $sign = 1$ and the initial phase difference was 0.25π, two oscillators came away from each other but did not synchronize with an anti-phase mode as shown in Fig. 5(d).

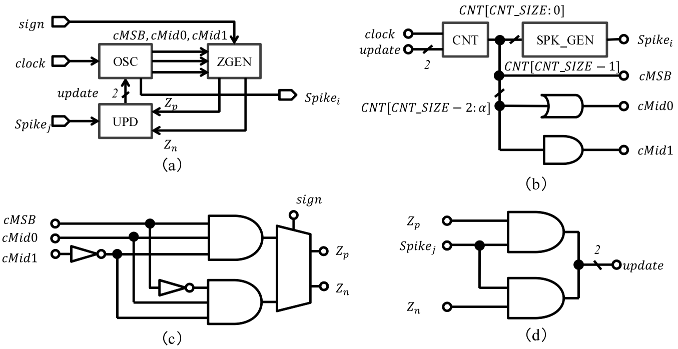


Fig. 3. Design of pulse-coupled phase oscillator circuit: (a) Top moduleC(b) Oscillator circuit (OSC)C(c) Function generator circuit (ZGEN)C(d) Update circuit (UPD)

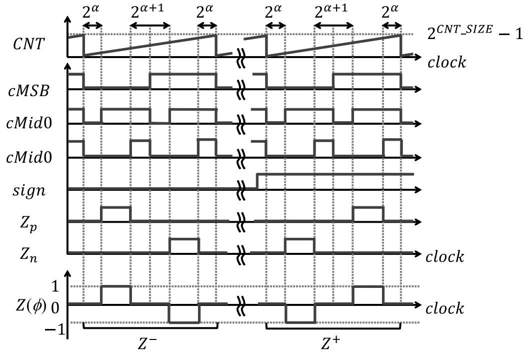


Fig. 4. Timing diagram for generating ZGEN

From the above simulation results, we confirmed that the proposed digital circuit emulated three kinds of synchronization phenomena.

4.2 Results of FPGA Implementation

The proposed circuit shown in Fig. 3 was synthesized by Precision logic synthesis tool [8]. The target FPGA device was Altera Stratix II EP2S60F672C [1].

Table 1 shows a synthesized result of the pulse-coupled phase oscillator circuits. Circuit parameters CNT_SIZE and α were the same as Section 4.1. These results show that our proposed circuit can be realized with small FPGA resources, and we expect that the proposed design approach enables to make a large pulse-coupled network emulator in a massively parallel manner.

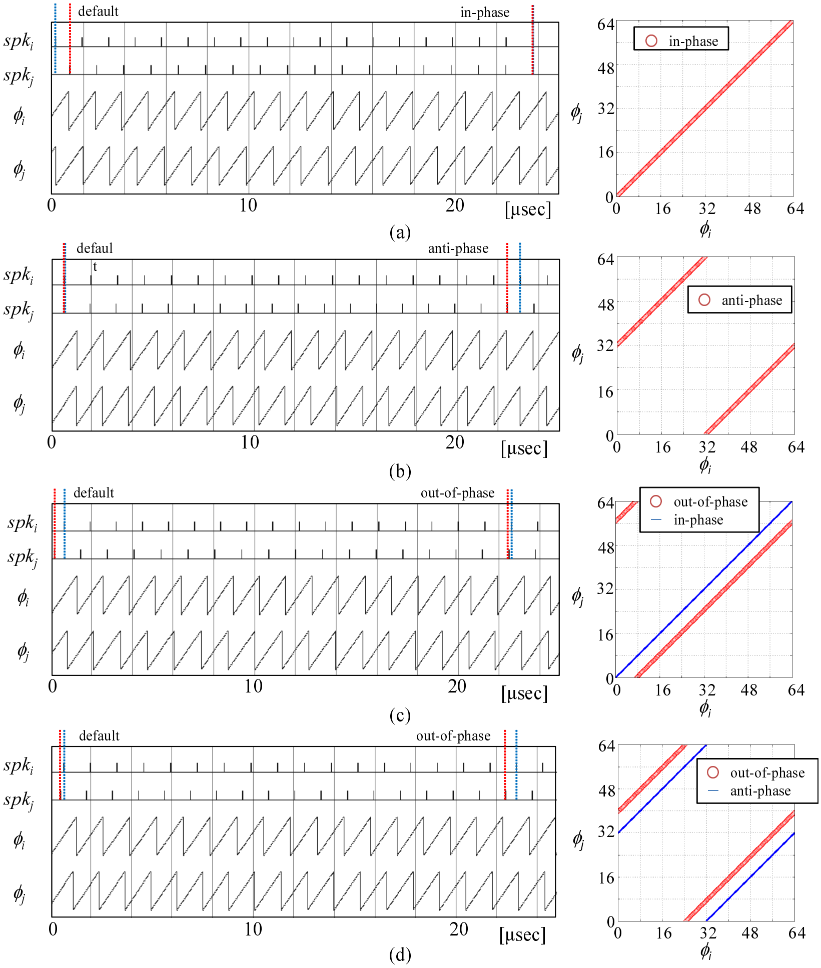


Fig. 5. Simulation results: (a) in-phase, (b) anti-phase, (c) out-of-phase(1), (d) out-of-phase(2)

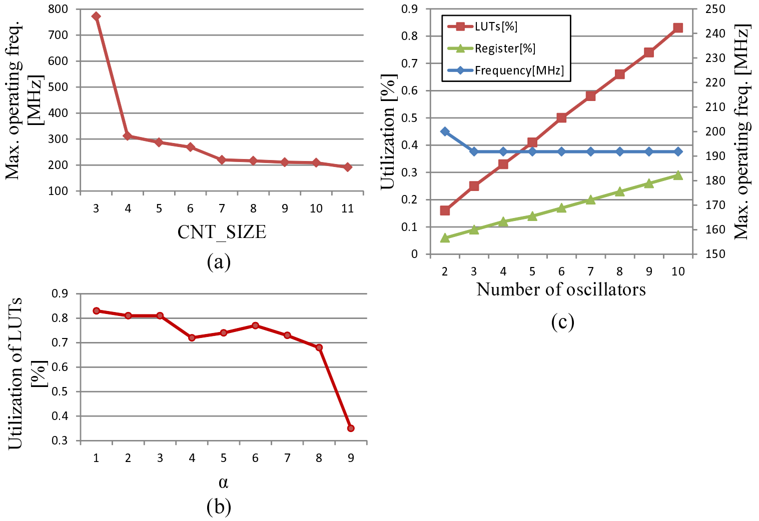
Figure 6 shows relationships between the circuit parameters and synthesized results. Maximum operation frequency f_{max} lowers with increasing CNT_SIZE, as shown in Fig. 6(a). It means that there is a trade-off between the two. Parameter α has a little effect on the utilization ratio of LUTs as long as $\alpha < CNT\_SIZE - 2$ ($CNT\_SIZE + 11$ in this case), as shown in Fig. 6(b).

Also, f_{max} is independent of the number of oscillators, as shown in Fig. 6(c). This means that the proposed approach enables a massively parallel circuit architecture.

Table 1. Resource and frequency reports

State	Resource	Used	Avail.	Utilization
in-phase anti-phase	LUTs	70	48352	0.14%
	Registers	26	48352	0.05%
out-of-phase	LUTs	70	48352	0.14%
	Registers	26	48352	0.05%

State	Frequency
in-phase	281.611MHz
anti-phase	
out-of-phase	538.793MHz



(a)

(b)

(c)

Fig. 6. Relationship between circuit parameters and synthesized results

5 Conclusion

We proposed a parameterized digital circuit design approach for pulse-coupled phase oscillators. Our approach aimed to construct a reconfigurable hardware platform that emulates a large-scaled analog pulse-coupled phase oscillator network. The proposed design achieves a small circuit size and high operating frequency because oscillators are coupled with only spike pulses and the operation part consists of only combinational circuits. The precision of calculation, the shape of coupling function and the number of oscillators in the network can be parameterized. Thus, it is easy to change the circuit architecture and to emulate various types of pulse-coupled networks. We observed three synchronization phenomena similar to those obtained in the analog LSI implementation. The results showed that the analog model was reproducible on the digital hardware model.

In future work, we will apply the proposed design to a coupled MRF model. We will extend the proposed parameterized design to a reconfigurable platform to realize a large-scaled coupled MRF network and its emulator. After we verify behavior of the coupled MRF network by the emulator, we will feed back its knowledge to ASIC implementation to realize an ultra-low power consumption neuromorphic VLSI, and then we will apply it to autonomous robot vision.

Acknowledgment. This research is partially supported by the Aihara Project, the FIRST program from JSPS, initiated by CSTP.

References

1. Altera Corporation, `http://www.altera.com/`
2. Blake, A., Zisserman, A.: Visual Reconstruction. MIT Press, Cambridge (1987)
3. Geman, D., Geman, S., Graffigne, C., Dong, P.: Boundary detection by constrained optimization. IEEE Trans. Pattern Analysis and Machine Intelligence 12(7), 609–628 (1990)
4. Kawashima, Y., Atuti, D., Nakada, K., Okada, M., Morie, T.: Coarse image region segmentation using region- and boundary-based coupled MRF models and their PWM VLSI implementation. In: Proc. Int. Joint Conf. on Neural Networks (IJCNN), pp. 1559–1565 (2009)
5. Lumsdaine, A., Waytt, J., Elfadel, I.: Nonlinear analog networks for image smoothing and segmentation. In: IEEE Proc. of Int. Symp. Circuits and Systems (ISCAS), pp. 987–991 (1990)
6. Matsuzaka, K., Morie, T.: A simplified region-based coupled MRF model for coarse image region segmentation toward its VLSI implementation. In: Proc. Int. Symp. on Nonlinear Theory and Its Applications (NOLTA), pp. 202–205 (2009)
7. Matsuzaka, K., Tohara, T., Nakada, K., Morie, T.: Analog CMOS circuit implementation of a pulse-coupled phase oscillator system and observation of synchronization phenomena. Nonlinear Theory and Its Applications, IEICE 3(2), 180–190 (2012)
8. Mentor Graphics Corporation, `http://www.mentor.com/`
9. Winfree, A.T.: The Geometry of Biological Time. Springer, New York (1980)

Brain-Inspired e-Learning System Using EEG

Kiyohisa Natsume

Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu-ku, Kitakyushu City, Fukuoka, 808-0196, Japan
natume@brain.kyutech.ac.jp

Abstract. English rhythm instruction materials (RIM) encourage one to learn English rhythm. In RIM, you have to speak loud following the English teacher's song looking at the phrases of the song in the text. During the learning the power of theta band (4 - 8Hz) of electroencephalogram (EEG) increased at the frontal region. When you repeat the lesson several times, you become bored. The powers of alpha (8-14 Hz), beta (14-30 Hz) and gamma (30-50 Hz) bands started to decrease in a wide regions before the subjects felt bored. On the other hand, theta power did not change. In addition, the coherence between the two recording sites mainly the electrode pairs along the midline was significantly different comparing between before and after subjects felt bored. The coherence of theta band did not change. These results suggest that using the characteristics of EEG, e-learning system for English rhythm can be developed.

Keywords: EEG, Boredom, α wave, β wave, γ wave, Coherence.

1 Introduction

Japanese has been widely accepted as a mora-timed language, while English is recognized as a stress-timed language. Nakano designed English rhythm instruction material (RIM) modified from Jazz Chants for Children [5], which includes audio and text with rhythmic symbols. The test group of students underwent a weekly 20 minute period of RIM-based instruction for one month period. Pre-test and post-test oral readings were recorded and the duration of the inter-stress interval (ISI) of each recording was measured. The results showed the students' inter-stress interval (ISI) was shortened following the RIM indicating the RIM is effective in enabling Japanese L2 learners to acquire English rhythm patterns. During learning RIM, the subjects' electroencephalograms (EEGs) were recorded. Theta power at the frontal part increased, and after finishing learning the power decreased [6].

We can see the word "boredom" in a dictionary. But the word has had few scientific basis. Two earlier studies have been reported using EEG. First is that the subjects have to push a button by their right or left hands for a long time, and then the coherence of β wave decreases [1, 4]. Second is that the subjects had to learn RIMs many times, and then β wave decreased [2]. We feel "bored" in various situations. For example, you feel bored when you are exercising, and then the exercising efficiency is down you feel. In the other situation, if you keep eating the same food, you feel bored

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 25–32, 2013.

and then you may dislike the food. Thus you use the word “bored” when you feel something bad. On the other hand, you can also assume the “bored” feeling as turning point of the behavior [3]. If we can scientifically verify whether you are bored or not, you can develop more efficient education system, know the timing you should have a rest in a lesson or exercise.

2 Materials and Methods

2.1 Recording Method for EEG

Eight healthy male subjects (average years 23.6 ± 0.53) took part in the present experiment. At first we recorded EEG when each subject opened his/her eyes and was calm for 30 seconds (control). After that, subjects had to have one of the lessons of the English RIM. In a lesson, they had to speak in a loud voice following the English teacher’s song given from the computer. They were instructed to raise their hands when they felt bored. The lyrics were shown on the monitor. They had to repeat the same RIM more than fifteen times. Total time was about 1000 sec. Eight EEG electrodes were put on the subject’s head according to the international 10/20 system all through the RIM lessons. EEG signal was amplified by the amplifier (X10000: DIGITEX LAB Co, Ltd), filtered between 0.5 and 100 Hz, and recorded in a computer using a LaBDAQ-2000 (Matsuyama Advance Co, Ltd) with the sampling rate at 1 kHz. Before starting this experiment they were instructed to raise their hands when they felt bored.

2.2 Analysis Method

In time-frequency analysis the time course of the spectrum powers was calculated by Fast Fourier Transformation (FFT) of the EEG signal using MATLAB software (Mathworks, Inc., USA). Time window of FFT is 500 msec, and the overlap time is 250 msec. The power at each frequency was divided into four wave groups, theta wave power (averaged from 4 to 8 Hz), alpha wave power (8 - 14 Hz) and beta wave power (14 - 30 Hz), and gamma wave power (30 – 50 Hz). The control power was averaged for 30 sec. They were averaged in the lesson per 70 sec, and subtracted the control power (Fig. 1). Coherence was calculated by the equation (1). The time window and overlap time are the same with FFT. S_{xx} and S_{yy} are auto spectrum of signal x or y. S_{xy} is cross spectrum of signal x and y (Fig. 1). Performance ratios are evaluated whether subjects could follow each rhythm in RIMs.

$$coh^2(\omega) = \frac{\left|S_{xy}(\omega)\right|^2}{S_{xx}(\omega)S_{yy}(\omega)} \qquad (1)$$

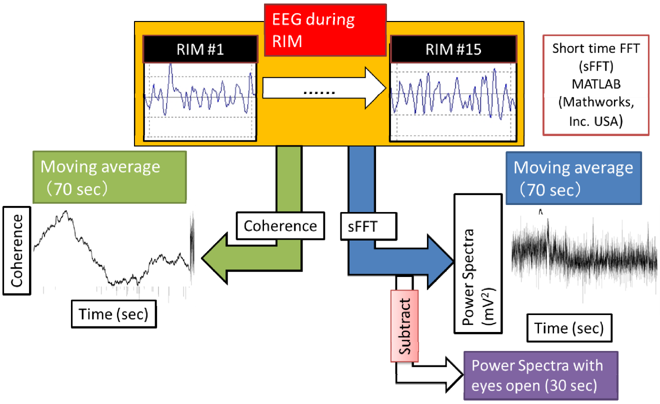


Fig. 1. The analytical method for the power and the coherence of θ, α, β, and γ waves. The time averaged control power was subtracted from the power at each frequency during RIMs.

3 Results

The typical temporal change of performance ratio in a subject is shown in Fig. 1. Six out of eight raised their hands after the ratio reached to 80%. The ratio of the rest of two was above 80% from the first. They raised their hands after a while after the start of the lesson

The power in the α, β, and γ bands remained constant near the start of the lesson, but it begin to decrease around 200 s after the start of a lesson at several electrode positions. After the decrease, the power values held constant. On the other hand, the power of θ wave remained constant not only just after the start, but also after >200 s. The subjects reported bored feeling after α, β, and γ power started to decrease (Fig. 3). Similar results were obtained across the eight electrode positions. The power of the α, β, and γ power waves decreased significantly at almost all electrode positions, while the power of θ wave did not decreased (Fig. 4).

We statistically compared the power before and after the subjects' raising their hands in Fig. 4. The three smallest significant probabilities were picked up and shown in Fig. 5. The probability of β wave at Oz is the smallest of all. The results suggest that among the power, the detection of decreasing in the power of β wave at central occipital area will be the most efficient to detect the bored state of the subjects.

Next, we calculated the coherences between the pairs of the recording positions. Temporal changes of coherence between Pz and Oz in subject A are shown in Fig. 6. Coherence also increased for the first time after starting the training, and after some while the coherence of the α, β, and γ waves started to decrease, while that of θ wave did not decreased. The decreasing coherence also reached to the steady state after some time. The subjects raised their hands after the time when the coherence of α, β, and γ waves started to decrease.

The coherence before and after raising the subjects' hand was statistically compared at all waves at all positions. The results show that the coherence at all bands except theta band decreased mainly on the electrode pairs along the midline. The three smallest significant probabilities were picked up and shown in Fig. 7. The probability of α wave at Fz-Oz is the smallest of all. There results suggest that among the coherences, the detection of decreasing in the coherence of α wave at central front-occipital area will be the most efficient to detect the bored state of the subjects. In addition, the order of the significant probabilities of the coherences are much lower than those of the powers.

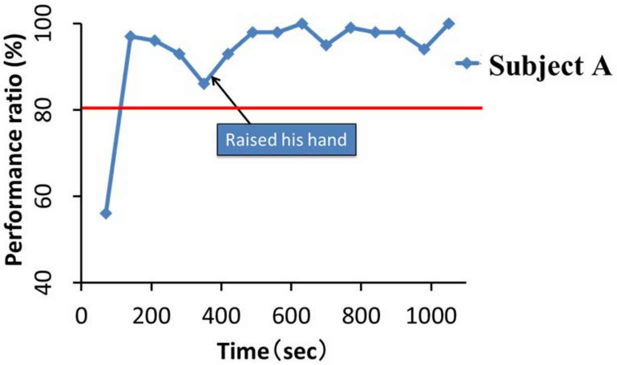


Fig. 2. The typical temporal change of performance ratio in a subject. Six out of eight raised their hands after the ratio reached to 80%. The ratios of the two were above 80% from the first.

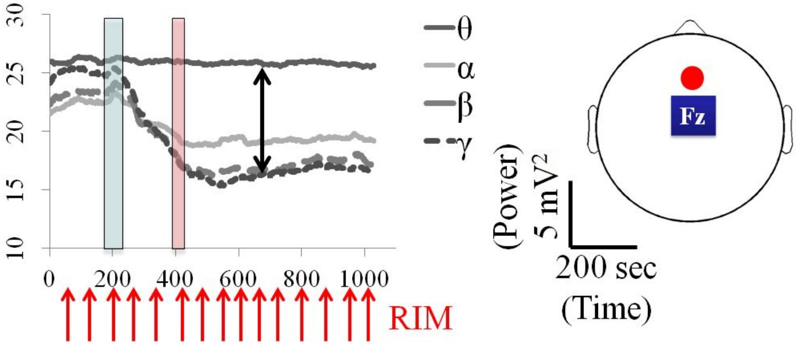


Fig. 3. Representative temporal changes in the powers of EEG in Fz location. The blue rectangle indicates the peaks of the power of each brain wave. The pink rectangle indicates the time of the subjects raising their hand. Time zero indicates the onset of the lesson in this and the following figures.

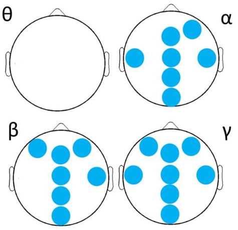


Fig. 4. The comparison of the power of each wave between before and after raising hands for 50 sec. Filled circles indicate the site where the power decreased and the significant probabilities are below 0.05 (Paired t-test; $p < 0.05$; $n = 6$).

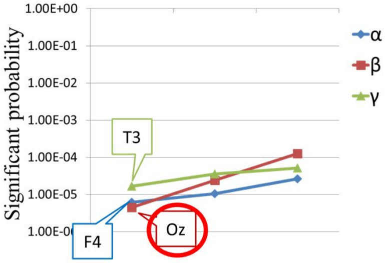


Fig. 5. The three smallest significant probabilities picked up from Fig. 4 are shown. The probability of β wave at central occipital area is the smallest of all (red circle). Significant probabilities are arranged in ascending order from left to right.

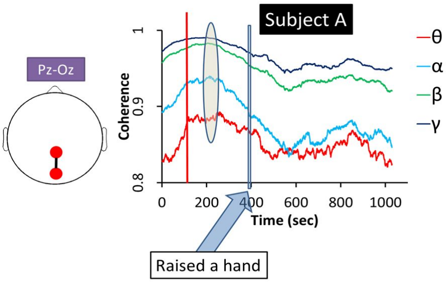


Fig. 6. Temporal change of coherence between Pz and Oz in subject A. The eclipse indicates the peak of the coherence of each brain wave. The thin blue rectangle indicates the timing of the subjects raising their hand. Red line indicates the time when the performance ratio of the subject reached to 80%. Coherence data are also analyzed with moving averaging method as Fig. 3.

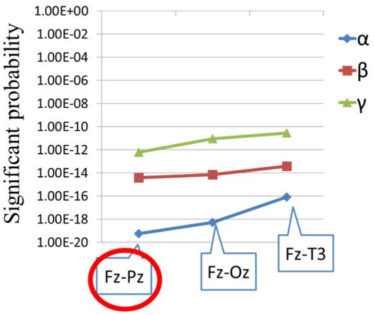


Fig. 7. The three smallest significant probabilities picked up from Fig. 6 are shown. The probability of the coherence of alpha wave recorded at Fz-Pz is the smallest (red circle). Significant probabilities are arranged in ascending order from left to right.

4 Discussion

In previous studies, researchers assumed that vigilance decrement is identical to boredom [4, 7]. The vigilance decrement has been described as a slowing of reaction times or an increase in error rates as an effect of time-on-task during tedious monitoring tasks. In those experiments, whether or not the subjects actually felt bored was not clarified. On the other hand, the present study elucidated the changes in EEG that occur when subjects felt bored. Therefore, EEG can be a better index for physiological "boredom" condition of a subject.

In the present study, after the power of α, β, and γ waves decreased widely across the skull, all subjects raised their hands. In contrast, the change in the power of θ wave was smaller than those of α, β, and γ waves. These results suggest that the decreases in α, β, and γ power across the head can cause the subjects' bored feelings. Because there are some delays between the start of the decrease in powers and the time when the subjects felt bored, it seems that it may take time for us to become consciously aware of the bored feeling, which results from unconscious neural activity.

The coherences of α, β, and γ waves also decreased mainly along the central front-occipital area, and all subjects felt bored. In contrast, the change in the coherence of θ wave was smaller than those of α, β, and γ waves. These results suggest that the decreases in α, β, and γ coherence can cause the subjects' bored feelings. Because there are some delays between the start of the decrease in coherences and the time when the subjects felt bored.

Which parameter can detect the subjects' bored feeling more efficiently, power or coherence? The averaged powers and coherence for 50 sec before and after the subjects felt bored were calculated, and compared. The significant probability of β at occipital region was the smallest among the comparison of powers, and the probability of α at front-occipital area was the smallest among the comparison of the coherence. The probability of the coherence had a smaller order than that of the power. These results suggest that we may easily detect the subjects' bored feeling using the coherence.

In the previous studies, decreased attention is thought to correspond with boredom [7]. Parieto-occipital α waves can contribute to attention [4]. In the present study, not only parietal α waves but also β and γ waves in other regions change with boredom. Boredom is accompanied by other processes than decreases in attention [7]; the change in brain waves across a wide area of the head—beyond parietal α waves—may reflect these processes. Further studies are necessary to confirm this possibility.

Other psychophysiological parameters than brain waves, heart rate, blood pressure, etc. [7] have been studied on the relationships with bored feelings. Compared with measuring heart rate or blood pressure, measuring EEG has a high time resolution. It would be very useful for the e-learning system to respond immediately when the learners feel bored.

As reported previously, after the powers of alpha, beta and gamma waves decrease, all subjects raised their hands. Coherence at most electrodes pairs decreased before raising their hands. These suggest that the decrease in the powers and coherence reflect subject's bored feeling. Because there are some delays between the start of the decrease in the decrease in power and coherence, and the timing of subjects raising their hands, it seems that it may take some time for us to realize our bored feeling after the response of EEG. In addition, performance ratio is not correlated with each subjects. Therefore, it's assumed using EEG powers and coherence is effective to identify boredom.

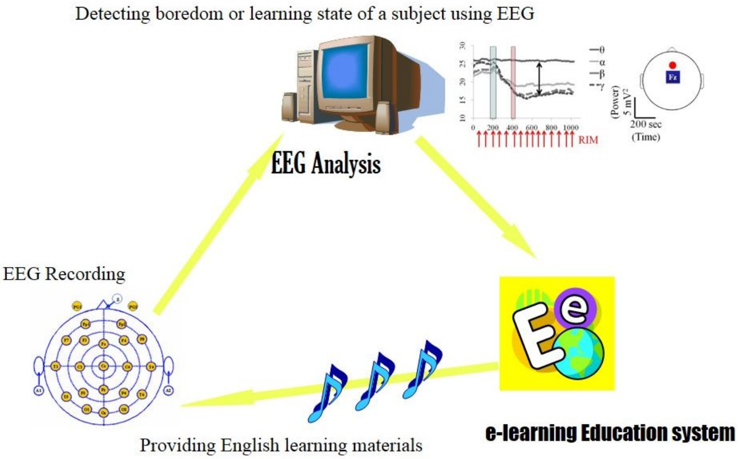

Detecting boredom or learning state of a subject using EEG
10
15
20
25
30
0
200
400
600
800
1000
RIM
θ
α
β
γ
(Power)
5 mV²
200 sec
(Time)
Fz
EEG Analysis
EEG Recording
Providing English learning materials
e-learning Education system

Fig. 8. Proposed e-learning system for learning English rhythm

Finally we propose e-learning system for learning English rhythm (Fig. 7). A learner's provided one of English learning materials via internet, and learn it while his/her EEG is recorded. EEG analysis machine is analyzing the feature of EEG, for example, the power and the coherence of EEG. When the machine is detecting the increase in θ power, the learner is learning the material with concentrated power, and the system keep providing the materials. When the system detects that the feature of EEG which reflects the learner's bored feeling, the system exchanges the material to a new one to refresh the learner's brain.

5 Conclusions

1. The power of α, β, and γ waves first kept constant after RIM learning and began to decrease as each subject repeated a RIM lesson.
2. Performance ratios during RIMs reached to 80% then the subject felt bored.
3. All subjects had bored feelings after the power of α, β, and γ waves began to decrease. The change in θ wave power was smaller compared with the changes.
4. The coherence of α, β, and γ waves began to decrease mainly along fronto-occipital line except that of theta wave.
5. All subjects felt bored after the coherence of α, β, and γ waves started decreasing.
6. The power spectra of θ wave did not change significantly, and the coherence of it changed with wide variability.
7. E-learning system for learning English rhythm using EEG is proposed.

Acknowledgment. This work was supported by KAKENHI Grant-in-Aid for challenging Exploratory Research (#22650205; 2010-2011, #24650353, 2012-2013).

References

1. Aslanyan, E.V., Kiroy, V.N.: Electroencephalographic Evidence on the Strategies of Adaptation to the Factors of Monotony. The Spanish Journal of Psychology 12(1), 32–45 (2009)
2. Katayama, T., Natsume, K.: The Change in EEG When You are Bored. J. Signal Processing 16(6), 637–641 (2012)
3. Kawamoto, H.: Tiring Force. NHK Publishing Seikatsujin-Shinsyo (2010) (in Japanese)
4. Lorist, M.M., Bezdan, E., ten Caat, M., Span, M.M., Roerdink, J.B., Maurits, N.M.: The Influence of Mental Fatigue and Motivation on Neural Network Dynamics; An EEG Coherence Study. Brain Research 1270, 95–106 (2009)
5. Nakano, H.: The Effect of Rhythm Instruction on Production Ability of Japanese EFL Learners. Annual Review of English Language Education in Japan 8, 81–91 (1997)
6. Nakano, H., Natsume, K.: English Rhythm Learning and Changes in EEG Using RIM with Beat. Computer & Education 13, 88–93 (2011) (in Japanese)
7. Ninomiya, K., Tetsutani, N.: A Research on "Tired" Factor Analysis Using Gaze-line Information. IEICE Tech. Rep. 109, 107–112 (2009)

A Method to Deal with Prospective Risks at Home in Robotic Observations by Using a Brain-Inspired Model

David Chik[1], Gyanendra Nath Tripathi[1], and Hiroaki Wagatsuma[1,2]

[1] Department of Brain Science and Engineering, Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu-Ku, Kitakyushu 808-0196, Japan
[2] RIKEN BSI, Japan
{chik-david,waga}@brain.kyutech.ac.jp,
tripathi-gyanendra-nath@edu.brain.kyutech.ac.jp

Abstract. Home robotics is a continuously growing field in academic research as well as commercial market. People are becoming more interested in advanced intelligent robots that can do housework and take care of children and elderly. A brain-inspired intelligent system is a possible solution to make the robot capable of learning and predicting risks at home. In order to solve difficult problems such as ambiguous situations and unclear causality, we propose a robotic system inspired from human working memory functions, which consists of an Event Map for storing observed information, and a Causality Map for representing causal relationships through supervised learning. The two maps couple together to enable the robot to evaluate various situations based on the appropriate context. More importantly, the Causality Map takes into account the dynamical aspects of physical attributes (e.g. the decreasing temperature of a hot pot). Our case studies showed that this is a satisfactory solution for predicting many risky situations at home.

Keywords: Home robotics, brain-inspired intelligent system, risk management, causality, learning.

1 Introduction

Recently, home robots such as automatic vacuum cleaners and robotic pets have been gaining popularity. People are expecting more advanced intelligent robots for doing housework and security check. The prospect is optimistic as Bill Gates believes that every home will have a robot [1], and South Korean Government stated that this popular home robot will be available by 2020 [2]. An important application of home robot is to look after elderly and children. In Japan, it is expected that the percentage of population over 65 years of age will increase to nearly 30% after 2020 [3]. In USA, there are nearly 2800 children die each year due to home accidents, which counts for 55% of all child deaths [4]. Indeed, looking after children and elderly requires a lot of human efforts. Therefore, it will be very useful if an intelligent robot is available for monitoring the home situation, providing a reliable evaluation of the safety / risk level, and then giving an appropriate response.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 33–40, 2013.

Learning and prediction of risks at home is a challenging problem. In particular, for some home accidents, there exists a flexible time lag between the cause and the effect (e.g. a knife was left on the table and then after a flexible time lag, a child reaches the table). This kind of situation cannot be analysed by conventional methods such as Granger causality which assumes a stable covariance within some time windows [5]. Probabilistic methods such as Bayesian inference [6] is also not appropriate in the case of home safety evaluation. For example, when a child is near a dining table, it is not useful to say that there is a 70% chance that the child will be safe and 30% chance that the child will be in danger. The robot must identify the exact causality in order to make useful prediction.

In this paper, we develop a new method for a robot to learn and predict what is dangerous and what is safe to a child by observing a teacher (e.g. a mother). This method is inspired from human brain mechanism. The hippocampus manages episodic memories to encode past behavioural temporal events with emotional expressions in accordance with the amygdala functions [7] and the prefrontal cortex plays a prominent role in decision making, which is involved in episodic memories in the hippocampus as contextual information [8]. In this sense, the prefrontal cortex is considered to maintain a function of logical reasoning by not only referring to semantic memory but also episodic memory being coupled together [9,10]. Based on this, we develop a robotic brain system which consists of an Event Map and a Causality Map.

2 The Robotic Brain System for Learning and Prediction of Risks at Home

2.1 Problem Formulation

Suppose there is a robot observing and recording some events that happened in the home environment. The goal of the robotic brain is to predict the consequences of events, and identify if some of the predicted events are dangerous. Mathematically, the problem can be formulated in this way:

$$G:\{E_i(x,y,t)\}(i=1,\cdots,k)\rightarrow\{E'_j(x,y,t)\}(j=1,\cdots,m) \tag{1}$$

where $\{E_i\}$ represents some previous and present events located in a 2D space at position (x,y) and time t (from some previous time $t=t_{now}-\Delta t$ to the present time $t=t_{now}$), as recorded by the robot; G denotes the reasoning process which maps some previous events (as causes) to some future events (as effects); and $\{E'_j\}$ represents the events in future at position (x,y) and some future time t as predicted by the robot. Some subsets of events in $\{E'_j\}$ may be dangerous (Fig. 1). In theory, it is difficult to construct Fig. 1. According to Frame problem [11] and Butterfly effect [12], an initial action causes changes in a large number of event-primitives in the real world. In practice, however, we assume that after the robot learns from human experts, the prediction of risky events is sufficiently reliable. Based on Eq. 1, we formulate a brain-inspired model, which consists of an Event Map for representing events ($\{E_i\}$ and $\{E'_j\}$) and a Causality Map for defining the knowledge of causality (G). Details of Event Map and Causality Map are explained below.

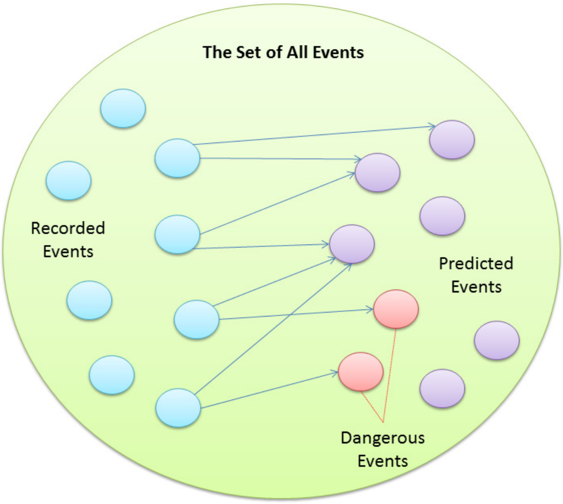


Fig. 1. Problem formulation. We consider a set of all events that can happen at home. The robot observes and records some events (blue circles). After that, it uses a reasoning process (represented as arrows in the figure). This reasoning process gives prediction to some future events (purple and red circles). A subset of the predicted events can be dangerous (red circles). The problem is: what is the best way for the robot to learn the events and the causality?

2.2 Event Map – Internal Representation of the External World

When the robot observes some events in the home environment, we assume that the observed information would be decomposed into many event-primitives. This is based on the fact that human brain also decomposes an object into different functional features [13]. Event-primitives include the position, time, states and physical attributes of an object. They are organized in a hierarchical structure, as shown in Fig. 2. We assume that there are 3 classes of objects: people/animals; movable; and fixed. Examples of people/animals include child, mother, dog, cat, etc. Examples of movable objects include book, toy, pot, knife, etc. Examples of fixed objects include bathtub, dining table, bed, sofa, etc. For each object, the robot keeps track of the states (how people can treat the object) and attributes (physical properties of the object). Some examples have been provided in Fig. 2.

The Event Map is used to store the position and time of events observed by the robot. From Eq. 1, we define an event-primitive $\{E_i(x,y,t)\}$ which is the value of a state or physical attribute of an object in the 2D space (x,y) and time t. Hence, an event is defined as a change of state or physical attribute in the space-time. The values of $\{E_i(x,y,t)\}$ may be Boolean or real numbers being normalized within 0 and 1. For example, whether a knife is being sheathed or not can be described as 1 or 0; the temperature of water can be described as a real number by normalizing between 0 and 100 degree Celsius; etc. These values are generated by observation of the robot.

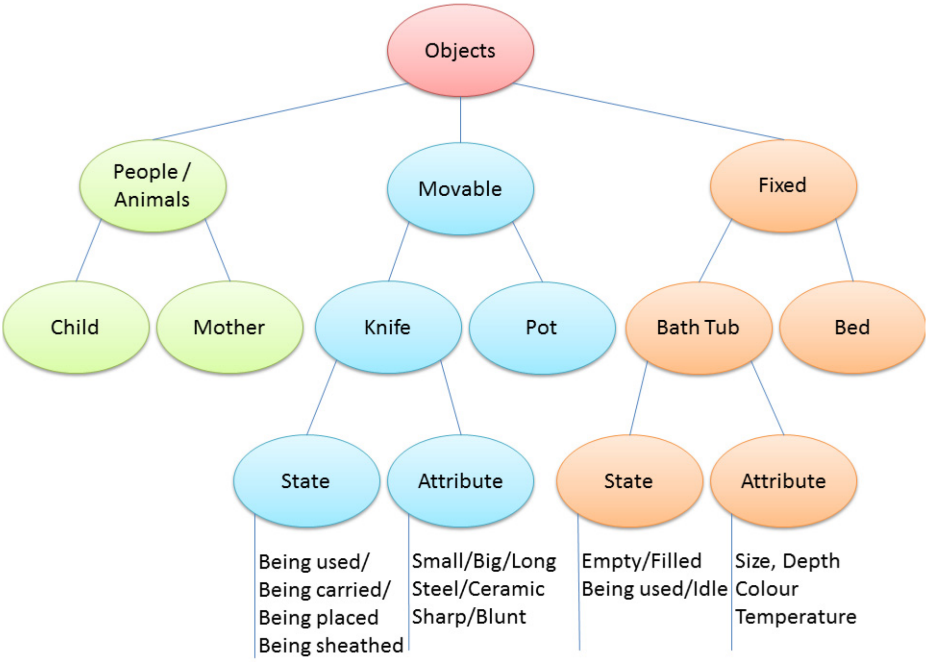


Fig. 2. The robot has an ontological understanding on objects and their physical properties. Values (in Boolean or normalized real number) of the states and attributes will be used as event-primitives in the Event Map

2.3 Causality Map – Learning the Knowledge of Risks at Home

The Causality map is a neural network-type system that learns and stores the knowledge of causality. The Causality map consists of two learning mechanisms: synaptic connections and intrinsic dynamics, which is recently discovered in neurons known as "intrinsic excitability" [14]. In our model, a future event E_1' may occur due to occurrences of a combination of multiple previous events E_1, E_2, E_3 as shown in Fig. 3. As a similarity to the synaptic plasticity, a weight $w_{E_i,E_i'}$ can be considered to change the relationship between events, according to the learning.

Secondly, the intrinsic dynamics represents a dynamics of intrinsic property or value, as we mentioned event-primitives in the Event Map (Section 2.2). Important point is that the value changes according to time, having specific time profile such as a flat, decaying, growing, and Gaussian shaped function depending on the internal dynamics. For example, a decreasing temperature of a hot pot can be modeled by a decay function of the form $E_{pot.temp} = e^{-\gamma(t-t_0)}$ according to Newton's law of cooling, while the occurrence of a child playing with a toy can be modeled by a Gaussian function of the form $E_{child.play} = e^{-(t-t_0-t_{peak})^2/2t_{width}^2}$ according to behavioral modeling [15], where t_0 is the onset time of the event. Based on the Causality Map, the robot makes some predictions of future events, which project back to the Event Map: $\{E_i\} \rightarrow \{E'_j\}$. If the predicted events are risky, then the robot will make an alert signal.

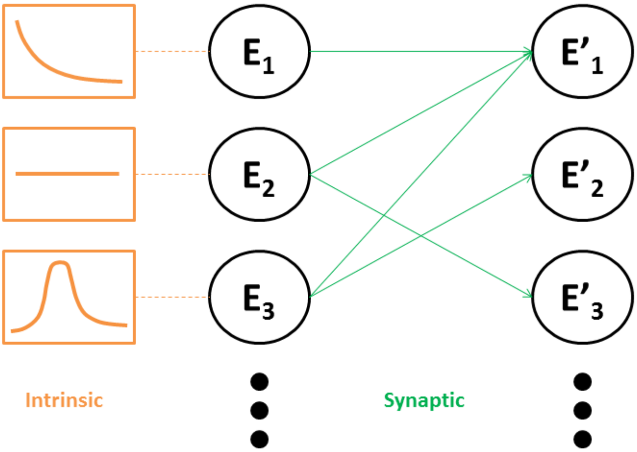


Fig. 3. Basic design of Causality Map, which consists of intrinsic dynamics for describing the internal changes of event-primitives with time, and synaptic connections for describing the interactions between event-primitives

3 Case Studies

We provide some examples below to illustrate how Causality Map can represent different kinds of dynamical causal relationship between events.

3.1 Child and Bookshelves

Let us consider a child walking towards the bookshelves. There are many possible future events: the child may take a book and read it; the child may climb up the bookshelves for fun (but dangerous); or the child may just leave and go to another place; etc. These possibilities are represented by connecting some arrows from event-primitive child.position(x,y) = bookshelves.position(x,y) to some possible event-primitives such as child.state = reading_book; child.state = climbing; etc. (Fig. 4). One of the possible events (child.state = climbing) is risky. Therefore, the robot will monitor the child and if necessary, prevent her from climbing.

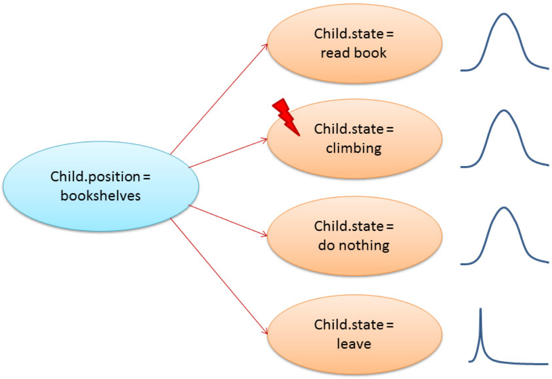


Fig. 4. An example showing the present state (the child walks towards the bookshelves) in blue circle; and some possible future events in orange circles

3.2 Toy in Plastic Bag

Suppose the child receives a present which is a toy inside a plastic bag. There are again many possible future events. The child may play with the toy; or the child may play with the plastic bag (which is dangerous due to the risk of suffocation). In this case, the robot considers two present states (child.position(x,y) = toy.position(x,y) and child.position(x,y) = plastic_bag.position(x,y)). The two present states project some possible future events as shown in Fig. 5.

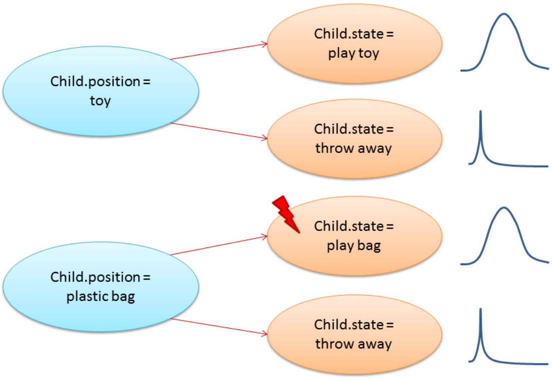


Fig. 5. An example showing two present states (the child is near a toy which is inside a plastic bag) in blue circles; and some possible future events in orange circles

3.3 Temperature of Pot

Suppose the child walks towards a pot. There is a possibility that the child may touch the pot in future. However, whether it is dangerous or not depends on the temperature of the pot. In this case, the robot considers pot.attribute = temperature and uses this information to decide whether the consequence is risky or not (Fig. 6).

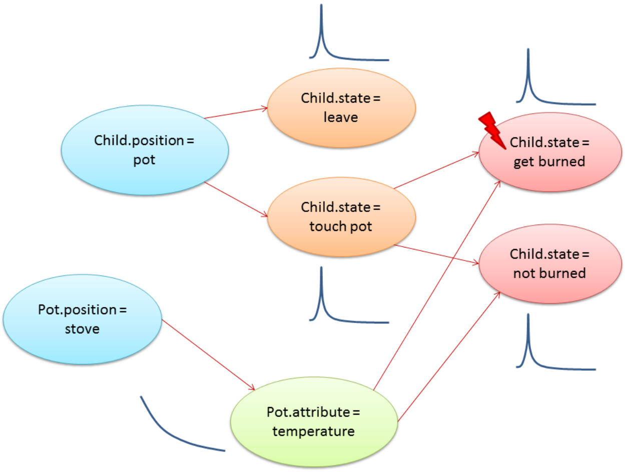


Fig. 6. An example showing the present state (blue circle) and future states (orange circles) of a child. In addition, the physical property of a pot (green circle) is taken into account to evaluate the risky level of consequences (red circles).

4 Conclusion

In this paper, we introduced a theoretical framework of a brain-inspired model for a robot to learn and predict risks at home. The system consists of an Event Map for representing events happening in the home environment, and a Causality Map for representing the dynamics and causality between events. We provided some examples to show that this design can encode many different kinds of causal and dynamical patterns. A rule based fuzzy cognitive map has been proposed to learn causal relations [16]. In their approach, membership functions are used to provide a flexible combination of connections with different degrees of change, which can be incorporated into the formulation of synaptic connection in our model. However, the fuzzy cognitive map does not consider changes of individual properties that happen internally. In our brain-inspired model, an important idea is that both synaptic and intrinsic parameters are considered. This provides multi-dimensional flexibilities in representing causal relations according to the spatio-temporal context. One unsolved problem is how to provide a general form for the intrinsic dynamics and derive a unique learning rule. In addition, we shall also need to solve implementation problems such as the limitations of robotic observations by external sensors. We hope that by solving these problems, the important function of home safety evaluation will be achieved as soon as possible, which will become a substantial benefit for the society.

Acknowledgments. The authors also gratefully acknowledge helpful supports in the Brain-IS project in Dept. of Brain Science and Engineering, Kyushu Institute of Technology. This work is partially supported by JSPS KAKENHI (22300081) and JSPS KAKENHI (24650107).

References

1. Gates, B.: A robot in every home. Scientific American 296(1), 58–65 (2007)
2. Lovgren, S.: A Robot in Every Home by 2020, South Korea Says. National Geographic News (September 2006)
3. Kaneko, R., Ishikawa, A., Ishii, F., Sasai, T., Iwasawa, M., Mita, F., Moriizumi, R.: Population projections for Japan: 2006-2055, outline of results, methods, and assumptions. The Japanese Journal of Population 6, 76–114 (2008)
4. Nagaraja, J., Menkedick, J., Phelan, K.J., Ashley, P., Zhang, X.L., Lanphear, B.P.: Deaths from residential injuries in US children and adolescents, 1985-1997. Pediatrics 116, 454–461 (2005)
5. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3), 424–438 (1969)
6. Winkler, R.L.: Introduction to Bayesian Inference and Decision, 2nd edn. Probabilistic Publishing (2003)
7. LeDoux, J.: The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon & Schuster Ltd. (1996)
8. Janowsky, J.S., Shimamura, A.P., Squire, L.R.: Source memory impairment in patients with frontal lobe lesions. Neuropsychologia 27, 1043–1056 (1989)
9. Wunderlich, K., Beierholm, U.R., Bossaerts, P., O'Doherty, J.P.: The human prefrontal cortex mediates integration of potential causes behind observed outcomes. Journal of Neurophysiology 106, 1558–1569 (2011)
10. Buehner, M., Krumm, S., Pick, M.: Reasoning = working memory ≠ attention. Intelligence 33, 251–272 (2005)
11. McCarthy, J., Patrick, H.: Some Philosophical Problems from the Standpoint of Artificial Intelligence. Machine Intelligence 4, 463–502 (1969)
12. Hilborn, R.C.: Sea gulls, butterflies, and grasshoppers: A brief history of the butterfly effect in nonlinear dynamics. American Journal of Physics 72(4), 425–427 (2004)
13. Schyns, P.G., Goldstone, R.L., Thibaut, J.-P.: The development of features in object concepts. Behavioral and Brain Sciences 21, 1–54 (1998)
14. Zhang, W., Linden, D.J.: The other side of the engram: experience-driven changes in neuronal intrinsic excitability. Nature Review Neuroscience 4, 885–900 (2003)
15. Modis, T.: The normal, the natural, and the harmonic. Technological Forecasting and Social Change 74(3), 391–398 (2007)
16. Carvalho, J.P., Tomé, J.A.B.: Rule based fuzzy cognitive maps - fuzzy causal relations. In: CIMCA 1999 - Computational Intelligence for Modelling, Control and Automation, Viena, Austria (March 1999)

A Study on Region of Interest of a Selective Attention Based on Gestalt Principles

Hyunrae Jo, Amitash Ojha, and Minho Lee*

School of Electronics Engineering, Kyungpook National University,
1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, South Korea
hrjo@ee.knu.ac.kr, {amitashojha,mholee}@gmail.com

Abstract. We propose a computational model to extend the region of attention in a visual scene. We assume that the visual information that is collected through bottom-up process is integrated by various mechanisms of perception process which in result further decides the attention regions of the object to accurately determine the object. This cycle is known as perception-action cycle. In our study we try to quantify relation between initial attention region and surrounding regions using Gestalt principles.

Keywords: Visual attention, gestalt principles, action-perception cycle.

1 Introduction

A very common notion about human visual perception is that humans can perceive an image in their visual field at once. However, recent research suggests that human visual system is far more complex and various sub-processes are involved in a single case of visual perception. Recent studies have shown that fragments of a visual scene are focused or attended to in a scene perception. For example, Pettet and Gilbert (1992), in their study, found dynamic activation changes in receptive-field size in cat's primary visual cortex. They claimed that neurons in low-level areas of the visual cortex (V1) extract low-level features in their small receptive fields [1]. This indicates the role of attention, which plays an important role by continuously moving and focusing on various parts of the visual scene to extract feature information in order to understand the whole scene. In other words, first few attention induces actions (eye-movement, etc) and estimates the perception result. This partial perception in return guides attention and searches for relevant information from the scene to confirm and complete the perception process. This process in which action and perception mutually contributes to each other is commonly known as 'Action-Perception cycle' [2].

Based on the above stated assumption that visual scene is represented by a set of image fragments, our aim, in this research is to find meaningful regions or features from bottom-up information of image fragments (like short contour elements, small patches and so on) that attract human attention and consequently helps perception. In this regard, Psychologists have proposed various models to explain what features

* Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 41–48, 2013.

attract human attention and how human visual system organizes a complex visual input (having different kinds of features) into unitary object(s). Gestalt principle is one of those models, which explains the organization mechanism of features in humans [3]. In last few decades, several Gestalt principles have been suggested. Gestalt attempts to describe how people organize visual elements into groups or unified wholes when certain principles are applied.

In the field of computer vision, Gestalt principles are widely used in detection, restoration, and segmentation. Gestalt principles have been applied especially on segmentation method. For example, Koostra and Kragic (2011) used an additional variable, Gestalt principle, to measure the goodness of a segment [4]. Similarly, Richtsheld and his colleagues (2012) focused on finding relations between surface patches [5]. They used Gestalt principles to determine if patches belonged to same kind to form an object.

The latest research on attention uses bottom up information based on human visual system. Itti, Koch, and Niebur (1998) introduced a computational model of focal visual attention called the saliency map (SM) [6]. Saliency map is a brain-like vision model that imitates receptive parts of human visual cortex. It uses bottom-up features such as color, intensity, and edge information to infer salient regions and compare them with surrounding environment. Jeong and Lee (2008) in their research used entropy variation between different attention areas for finding meaningful boundaries of an object [7].

In this paper, we propose a method to expand initial attention area, based on perception-action cycle, by modeling three Gestalt principles namely (1) principle of similarity, (2) principle of continuity and (3) principle of proximity. We also explain our method of measuring these gestalt principles. Finally we present our experiment results.

2 Background

2.1 Perception-Action Cycle and Spread of Attention

An action (like eye movement) can spark from previous perception experiences, if the available information from a visual stimulus is insufficient to recognize it. Basically, the perception-action cycle is the circular flow of information that takes place between the organism and its environment in the course of a sensory-guided sequence of behavior towards a goal [2]. In our study we define “perception” as the gestalt relation between initial attention area and surrounding area. It is calculated by basic features in attention area such as color, edge and so on. The “action” is defined as shift of attention, the extent of its movement in image, and its size of window to get sufficient information. An indirect support for perception-action cycle comes from the study of Roelfsema and Houtkamp (2011), They proposed an incremental grouping theory [8]. This theory addresses the spread of enhanced neuronal activity that corresponds to the labeling of image elements with object-based attention. Fig. 1 shows an example of incremental grouping theory. In this example, there are two zebras having their own distinguishing features that have distinctive qualities from surrounding objects and background. For example, there are colors, contours and striped patterns. If we have an initial attention area (which is located in the object), then it can be compared with surrounding areas. The attention spreads until an object is represented as a perceptual group (whole). It is a meaningful region even if the whole object does not cover the attention boundary.

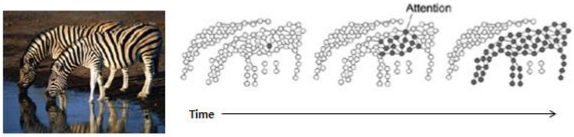


Fig. 1. Incremental grouping in natural image (Source: Roelfsema et al. (2011), [8])

2.2 Gestalt Principles

Gestalt principles explain how parts are grouped as a unified whole by applying various principles. In Gestalt theory, relations between elements are defined by several characteristics. For example, similarity, proximity, common fate, connectedness, good continuity, etc. These characteristics are also known as laws or principles of Gestalt. Several brain studies have confirmed the application of Gestalt principles. For instance, Wannig et al. (2011), recorded neuronal activity in area V1 of macaque monkeys and observed an automatic spread of attention to image elements outside of the attentional focus when they were bound to an attended stimulus by Gestalt criteria [9]. This result shows neurological correlates of operating Gestalt principles. Fig. 2 shows some images that are grouped differently using various principles of Gestalt. Humans are aware about which part belongs to a group explained by which principle.

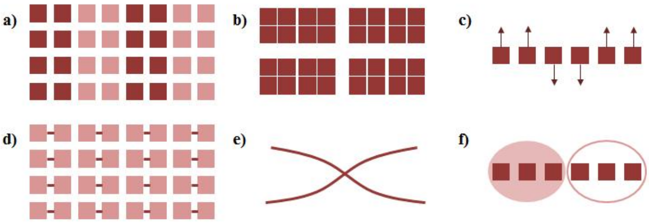


Fig. 2. Examples of Gestalt principles: (a) Similarity, (b) Proximity, (c) Common fate (d) Connectedness, (e) Good continuation, (f) Common region

3 Proposed Model

Roelfsema et al. [8] conjectured that Gestalt grouping is implemented by connecting neurons tuned to image features that are likely to belong to the same perceptual object. Koostra and Kragic used Markov random field and graph-cut techniques for segmenting the object from the background. They tried to quantify seven Gestalt principles to improve the accuracy of segmentation model.

Our study is partly motivated by Koostra and Kragic's attempt to quantify Gestalt principles in computer vision. In our model we focused on expansion of attention. We gradually expand the area of interest using Gestalt relation between visual elements. For implementing our model, we define an attention region with 30x30 window. It is an atomic element which can formulate relations with other regions (see Fig. 3 (b)). Initial attention point is given by maximum of the result of saliency map [6].

An attention region can consist of foreground and background, and distinguishing them is a problem. For example, consider the problem of background and foreground recognition in Rubin's cup problem (Fig. 3 (a)). In this optical illusion, people can not see two faces (with black as a foreground) and a cup (with white as a foreground) at the same time. To solve this problem, however, in our method, we select the object or foreground part using saliency map and assume that densest part is more likely to be an object.

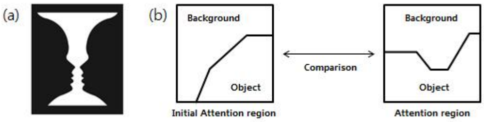


Fig. 3. (a) Rubin's cup, (b) The comparison between two attention regions

In our model, we consider three Gestalt principles for defining attention region namely (1) similarity, (2) good continuation and (3) proximity.

First, the principle of similarity is defined by the similarity of features. The color similarity is measured as:

$$G_c = 1 - \frac{\sqrt{|v_1|^2 + |v_2|^2 - 2|v_1||v_2|\cos\theta}}{\sqrt{|v_1|^2 + |v_2|^2 + 2|v_1||v_2|\cos\theta}} \tag{1}$$

$$|v_i| = \sqrt{{r_{oi}}^2 + {g_{oi}}^2 + {b_{oi}}^2} \tag{2}$$

$$\cos\theta = \frac{r_{o1}r_{o2} + g_{o1}g_{o2} + b_{o1}b_{o2}}{|v_1||v_2|} \tag{3}$$

$$G_{cc} = \sqrt{0.59(r_o - r_b)^2 + 0.3(g_o - g_b)^2 + 0.11(b_o - g_b)^2} \tag{4}$$

Color similarity G_c is calculated from RGB vector (v_i) between interest areas [10]. r_o, g_o, b_o is respectively mean of color in object region and have unit norm. Color contrast $G_{cc}$is defined as weighted distance in RGB color vector. r_b, g_b, b_b is respectively mean of color in background region and have unit norm.

The principles of good continuation is defined by the continuous and smooth contour of the object.

$$G_{gc} = \frac{1}{n}\sum\nolimits_{i \in C} \kappa \tag{5}$$

$$\kappa = \frac{\left|x' \, y'' \quad + y' \, x''\right|}{(x'^{2} + y'^{2})^{3/2}} \tag{6}$$

C in formula (5) is the set of all contour points. *N* is total number of contour points. *K* in formula (6) is curvature of contour.

And the principle of proximity is defined by Euclidean distance and grouping range is limited.

$$G_p = \sqrt{(x_i - x)^2 + (y_i - y)^2} \tag{7}$$

A relation between various attention regions can determine the extension of attention region. A visual representation of our proposed model is shown in Fig. 4.

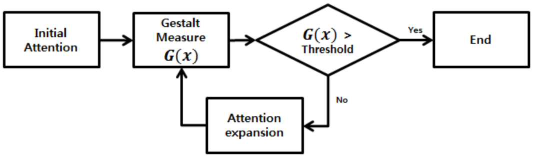


Fig. 4. A visual representation of attention model based on Gestalt principles

4 Results

To test our method we implemented it on several images. Fig. 5 shows the sequence of our model. First, initial attention region is acquired from saliency map. Initial attention region is the point of maximum density of saliency map and is compared with visual search process by parallel mechanism of attention in feature integration theory [11]. Perception occurs using several basic features such as color and contour of edge. Attention region is expanded by the relation of perception based on three Gestalt principles namely color similarity, continuity and proximity. Color similarity was used based on the assumption that an object is of one and same color. Continuity and proximity are used to restrict spreading direction and distance. The Gestalt relationship between initial attention region and surrounding region is expressed like a saliency map. We integrate those maps, and shift of attention is accomplished using this integrated map. At the end of this process, we can get a part of object. Next we present visual representation of our method as well as some of the results.

Fig. 6 explains the expansion of attention region by the iteration of the method in Fig. 5. Fig. 6 (a) is the original image and has two major locations of attention from saliency map that are specified from maximized region of saliency density. Fig. 6 (b) shows the relation image of color similarity, edge continuity and location proximity by Gestalt principles. When it is integrated, it generates Gestalt relation map and the shift of attention can be determined by Gestalt relation map. A cumulative attention by shift expands attention part. Fig. 6 (c) is the same process as with Fig. 6 (b) from the second location of initial attentions.

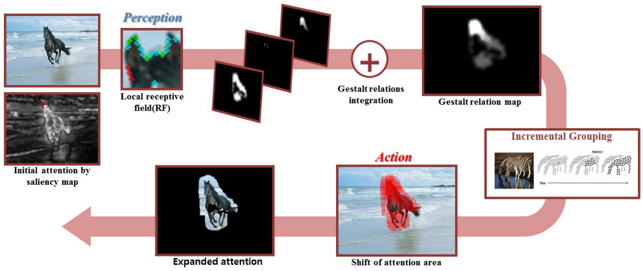


Fig. 5. The Sequence of our model

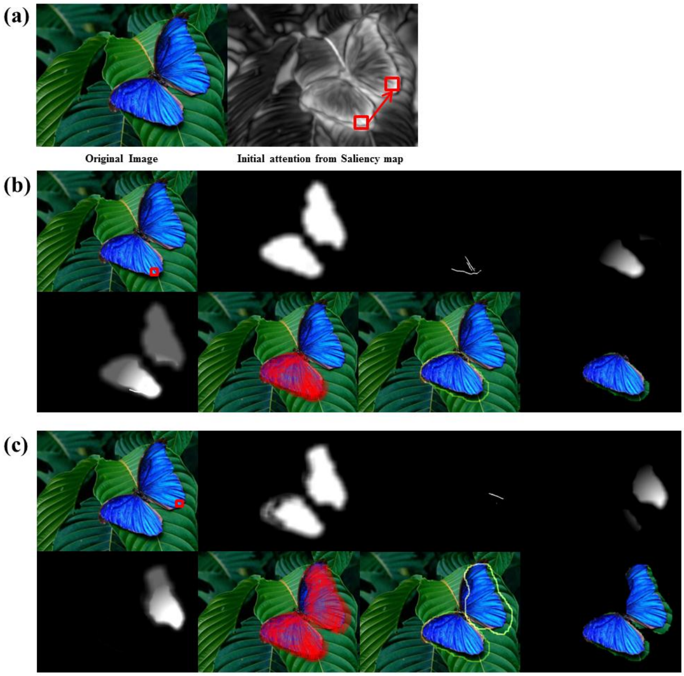


Fig. 6. The expansion of attention parts of an object using our method : a) Shift of attention by saliency map. We can get initial attention location from maximum density of saliency map. b), c) Spread of initial attention. It shows Initial attention, color similarity, continuity, proximity, integrated gestalt map, shift of attention, expanded boundary of attention, expanded region of attention. b) made from fist initial region, c) made from second initial region.

This process can be explained as 'Action-perception cycle' again. First, the perception occurs in initial attention area. Second, the area of attention is expanded using gestalt relations. This is the 'action process'. The combination of each attention part is used in recognition of the object (Perception). Fig. 7 shows the expansion area for several objects.

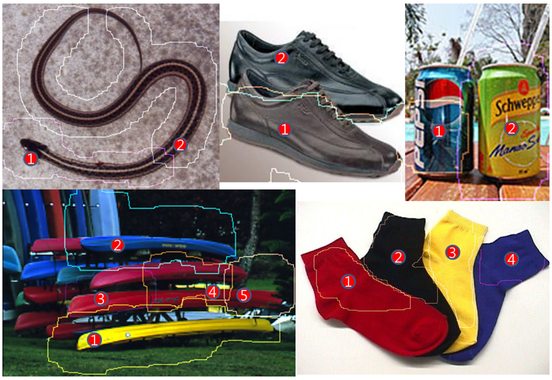


Fig. 7. Expanded region of attention

5 Conclusion and Future Works

In this paper, with the assumptions of perception-action cycle, we explored how perception result and information integration mechanism affect human visual attention. We tried to implement a computational model by quantifying Gestalt principles to determine the attention region in visual stimulus. We also explained our method of quantifying principles of similarity, continuity and proximity using primary information in visual system such as color, edge, and so on.

As a result, when an initial attention area is given, the result of perception, which is obtained by bottom-up information, determines next location of attention. The iteration of this process can define partial region of object.

Existing research on saliency map theory defines saliency region of a scene by feature integration with color and edge information. But in our study we achieve this goal of defining salient region by integrating Gestalt principles which also confirm to the assumption of "Action-Perception cycle". We assume that our approach is novel and closer to human visual perception mechanism and can generate better results in object recognition. .

In our future research, we plan to quantify more laws of Gestalt and selectively choose relevant principles in a visual scene. We also plan to analyze the relationship between the flow of the gaze and gestalt principles.

Acknowledgment. This work was supported by the Industrial Strategic Technology Development Program(10044009) (50%) funded by the Ministry of Knowledge Economy(MKE, Korea) and also supported the Converging Research Center Program through the Ministry of Science, ICT and Future Planning, Korea (2013K000333) (50%).

References

1. Pettet, M.W., Gilbert, C.D.: Dynamic changes in receptive-field size in cat primary visual cortex. Proceedings of the National Academy of Sciences 89(17), 8366–8370 (1992)
2. Cutsuridis, V. (ed.): Perception-Action Cycle: Models, Architectures, and Hardware, vol. 1. Springer (2011)
3. Wertheimer, M.: Untersuchungen zur Lehre von der Gestalt. II. Psychological Research 4(1), 301–350 (1923)
4. Kootstra, G., Kragic, D.: Fast and bottom-up object detection, segmentation, and evaluation using Gestalt principles. In: 2011 IEEE International Conference on Robotics and Automation (ICRA), pp. 3423–3428 (2011)
5. Richtsfeld, A., Zillich, M., Vincze, M.: Implementation of Gestalt principles for object segmentation. In: 21st International Conference on Pattern Recognition (ICPR), pp. 1330–1333 (2012)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
7. Jeong, S., Ban, S.-W., Lee, M.: Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. Neural Networks 21(10), 1420–1430 (2008)
8. Roelfsema, P.R., Houtkamp, R.: Incremental grouping of image elements in vision. Attention, Perception, & Psychophysics 73(8), 2542–2572 (2011)
9. Wannig, A., Stanisor, L., Roelfsema, P.R.: Automatic spread of attentional response modulation along Gestalt criteria in primary visual cortex. Nature Neuroscience 14(10), 1243–1244 (2011)
10. Plataniotis, K.N., Androutsos, D., Venetsanopoulos, A.N.: Adaptive fuzzy systems for multichannel signal processing. Proceedings of the IEEE 87(9), 1601–1622 (1999)
11. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–136 (1980)

Triggered Initiation of Retrograde Wave Propagation in a Cable of FitzHugh-Nagumo Cells

Katsumi Tateno

Department of Brain Science and Engineering, Kyushu Institute of Technology,
2-4 Hibikino, Wakamatsu-ku, Kitakyushu, Japan
tateno@brain.kyutech.ac.jp

Abstract. We studied a triggered initiation of retrograde propagation of an action potential in a one-dimensional cable of FitzHugh-Nagumo models. The triggered events occurred in the phase-dependent manner. A single stimulation induced alternation in action potential duration at the peripheral part of the pacemaker elements. The retrograde propagation automatically stopped.

Keywords: Triggered event, FitzHugh-Nagumo model, Wave propagation.

1 Introduction

Complex patterns of excitatory wave, such as spiral wave [1] or rotors [2], appear in an excitable media. An abnormal wave propagation may lead to undesired results, such as cardiac arrhythmia. Such abnormal activity can be initiated at an ectopic focus or spatial inhomogeneity in the electrical properties of an excitable media. The abnormal activity paroxysmally initiates and stops.

The premature beat is an occasional premature depolarization in the heart. In the non-sustained premature ventricular contraction, two or more premature beats in a row appear between normal cardiac beats. Those abnormal beats automatically stop. Rapid excitatory event may introduce a fatigue in a cell. If the fatigue is accumulated, an abnormal activity is terminated.

Beat-to-beat changes in the shape of action potential has been considered to associate cardiac arrhythmia [3]. The shortening of the action potential duration (APD) allows to sustain the spiral wave. A beat-to-beat alternation may lead to instability of the spiral wave [4]. A cascade of bifurcations of the wave propagation can result in an unstable pattern of the cardiac electrical activity [5].

An electrical stimulation leads to dynamical changes in an excitatory wave propagation. In a model study, a single stimulation induces sustained repetitive activity in a one-dimensional cable of excitable cells [6]. In their model, ectopic pacemaker cells were assumed. The sustained retrograde propagation is triggered by a single stimulation. Another stimulation is needed for the termination of the retrograde wave propagation.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 49–54, 2013.

In the present study, pacemaker elements were located at the end of a one-dimensional cable of the excitable elements. The excitatory wave evoked by the pacemaker elements was bound in the neighbors. A single stimulation triggered the retrograde wave propagation in the phase-dependent manner. The retrograde wave propagation automatically stopped.

2 Methods

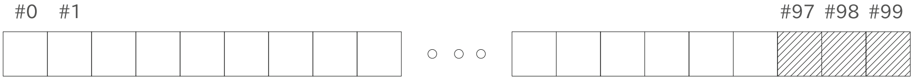


Fig. 1. A one-dimensional cable of the FitzHugh-Nagumo models. The cable consists of 100 FHN elements. The right 3 elements are periodically fired by a constant depolarizing bias current. The left 10 elements are stimulated by a short depolarizing pulse.

The cell model was the FitzHugh-Nagumo (FHN) equations. A one-dimensional cable of the FHN model was described as follows:

$$\frac{\partial v}{\partial t} = D\frac{\partial^2 v}{\partial x^2} + v(1-v)(v-a)w + I \tag{1}$$

$$\frac{\partial w}{\partial t} = \epsilon(v-w) \tag{2}$$

where v is the fast variable and w is the slow variable. D was a diffusion constant and I is a constant bias current. The parameters were $a = 0.139$, $\epsilon = 0.005$, and $D/(\Delta x)^2 = 0.038$.

A one-dimensional cable consisted of 100 FHN elements (Fig. 1). The right end of the cable was assumed to be abnormal pacemaker elements (#97 - #99). The pacemaker elements were depolarized ($I = 0.08$), and consequently spontaneous firing occurred. Other elements were excitable, but not spontaneously active ($I = 0$).

A single shot was given 10 elements at the left end of the cable. The duration of the stimulus pulse was 10(= 1000 steps). The amplitude of the stimulus pulse was 2.

An action potential was repeatedly generated at the pacemaker elements. When $D/(\Delta x)^2$ was below 0.042, the propagation of an action potential was bound in the peripheral part of the pacemaker elements. However, when $D/(\Delta x)^2 \geq 0.036$, an action potential evoked by a short depolarizing pulse propagated from the left end of the cable to the right end.

The element #91 was near the edge of the peripheral part of the pacemaker elements. The element #91 was forced to fire due to the pacemaking elements.

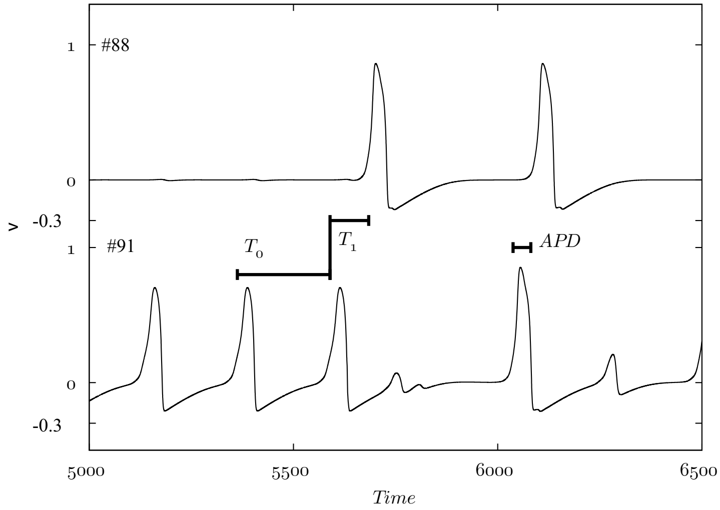


Fig. 2. Definition of the phase. The upper trace is the fast variable v of the element #88. The lower trace is the fast variable v of the element #91. The phase of the simulation ϕ is T_1/T_0.

Figure 2 shows the definition of the phase of the stimulation. T_0 was the interspike interval immediate before the arrival of the stimulation at the element #91. T_1 was the spike interval of the element #88 since the last action potential of the element #91. The phase ϕ of the stimulation was defined as T_1/T_0. The stimulation might elicit an action potential at the element #91 or lengthen the interspike interval. We measured the action potential duration (APD) after the stimulation. The APD defined as the time from 0.2 on the upstroke of an action potential to 0.2 on the repolarizing curve of action potential. Before the stimulation, the mean APD of the element #91 was 40.9 ms.

A explicit Eular scheme was used for the numerical integration. The time step (Δt) was 0.01. Neumann boundary conditions were used.

3 Results

Wave propagation induced by the pacemaker elements was limited in the peripheral part of the pacemaker elements. The excitatory wave did not reach the left end of the cable. The stimulation to the left 10 elements elicited an excitatory wave (Fig. 3a). The wave front reached near the pacemaker elements at $\phi = 0.33$. This excitatory wave did not induced retrograde wave propagation from the pacemaker elements (Fig. 3a).

When the timing of the stimulation was slightly delayed (= 2000 steps delay), the excitatory wave evoked by the stimulation induced the retrograde wave propagation (Fig. 3b). The retrograde wave propagation occurred after the arrival of

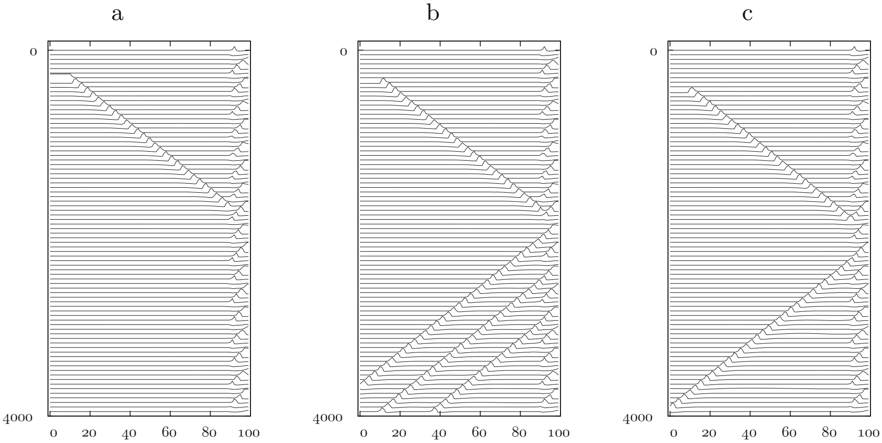


Fig. 3. Triggered initiation and spontaneous termination of the retrograde wave propagation. The horizontal axis is the cell number. The vertical axis is time. The membrane variable v was plotted every 5000 steps. a) $\phi = 0.33$. The stimulation does not induce the triggered wave propagation. b) $\phi = 0.42$. The action potential retrogradely propagates from the right end of the cable to the left end. The retrograde wave propagation automatically stops after three action potentials. c) $\phi = 0.9$. The retrograde wave propagation is not repeated.

the stimulation with a delay. A single stimulation triggered the wave propagation from the pacemaker elements. Three action potentials in a row appeared. The retrograde wave propagation automatically stopped.

The number of the triggered events depended on the timing of the simulation. In the early phase ($\phi = 0.12 \sim 0.38$), the triggered event did not occur. In the middle phase of the period ($\phi = 0.42 \sim 0.50$), 3 action potentials were allowed to propagate retrogradely. In the late phase ($\phi = 0.55 \sim 0.99$), 1 or 2 triggered action potentials occurred (Fig. 3c).

Figure 4 shows the trace of the fast variable v of the peripheral part (#90 and #91) and the pacemaker element #99. The stimulation failed to elicit an action potential of the element #91. After a long pause of the failure of excitation, a wide (APD = 43.4 ms) and high amplitude action potential appeared. The stimulation lengthened the interspike interval. The wide action potential of the element #91 elicited an action potential of the neighbor elements retrogradely. The retrograde wave propagation was initiated. The wide action potential was followed by a small depolarization (APD = 6.0 ms). The small depolarization did not evoke an action potential of the neighbors. A second wide action potential followed the small depolarization. The alternation of APD appeared several times. The APD alternation eventually ceased. The APD returned to the stable duration. Consequently, the retrograde propagation stopped.

When the stimulation arrived in the late phase of the period, the wide action potential occurred immediately after the stimulation. However, the first wide

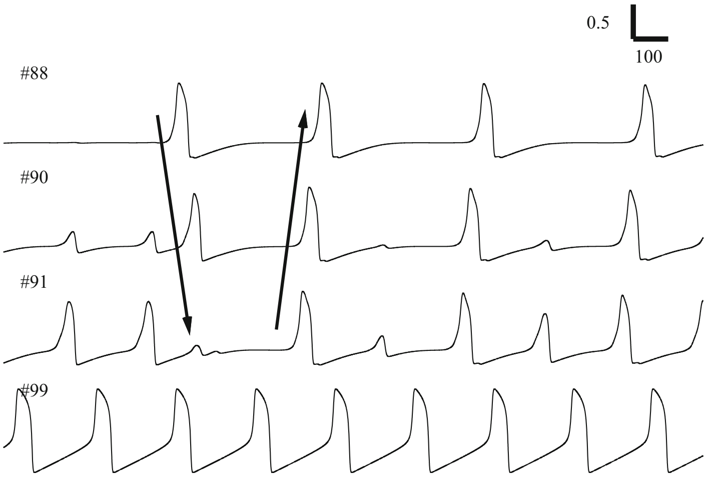


Fig. 4. The retrograde wave propagation. The stimulation (#88) elicits an action potential of the element #91. The large or small action potentials alternatively occur at the element#91. The alternation in action potential duration lasts several times after the stimulation. The thick arrow indicates the direction of the wave propagation.

action potential was blocked by the refractory period of the element #90. The second wide action potential was allowed to elicit an action potential of the neighbors. As the first wide APD did not contribute the retrograde propagation, the number of the triggered events reduced.

When the stimulation arrived in the early phase, the stimulation failed to elicit an action potential at the element #90. This is because the wave front of the simulation met a refractory period of the element #90. The stimulation was blocked before the element #91.

4 Discussion

The present study shows the triggered initiation and automatic termination of the retrograde wave propagation. A single stimulation at a critical phase initiates repetitive retrograde wave propagation. The wave front with a wide APD sequentially evokes an action potential of the neighbors. As the action potential duration returns to the stable value with time, the retrograde wave propagation was automatically terminated.

The triggered retrograde wave propagation depends on the phase resetting property of the peripheral part of the pacemaker elements. As the peripheral part is driven by the pacemaker elements, the periodic oscillation occurs. When the stimulation arrives in the middle phase of the period, the interspike interval is lengthened. The long interspike interval escapes the refractory period of the neighbor. Therefore, the retrograde wave propagation is drawn from the pacemaker elements. The stimulation at the late phase shortens the interspike interval. Such immediate response is blocked by the refractory period of the neighbor. This reduces the number of the triggered events.

A one-dimensional cable model proposed by van Capelle and Durrer [6] possesses the bistability: repetitive firing and quiescent. The cable is initially stable, but the stimulation starts periodic firing. The bistability provides a mechanism for the triggered activity of a network model of cardiac cells. In their network model, arrhythmia does not stop automatically. The termination of arrhythmia needs an extra stimulation. A fatigue term is necessarily introduced into the cell model for automatic termination of the retrograde propagation. The fatigue term reduces the excitability of the model cells during firing. Bub and his coworkers have reported bursting calcium rotors in cultured cardiac monolayer [7]. The calcium rotors paroxysmally start and stop. The triggered activity is reproduced by introducing spatial heterogeneity in their model [8]. Their cell contains a fatigue term. The present cable of the FHN models automatically stops after several wave propagation without a fatigue term.

A single stimulation draws the retrograde wave propagation out from the caged pacemaker activity in a one-dimensional cable of the excitable elements. The retrograde wave propagation automatically stops after two or three action potentials. Those features of the present cable are potentially a model for the non-sustained premature ventricular contraction. The number of the triggered events is determined by the adaptation time of the APD alternation. If the APD alternation lasts, the retrograde wave propagation turns into the persistent activity.

References

1. Winfree, A.T.: Varieties of spiral wave behavior: An experimentalist's approach to the theory of excitable media. Chaos 1, 303–334 (1991)
2. Winfree, A.T.: Electrical instability in cardiac muscle: Phase singularities and rotors. J. Theor. Biol. 138, 353–405 (1989)
3. Surawicz, B., Fisch, C.: Cardiac alternans: Diverse mechanisms and clinical manifestations. J. Am. Coll. Cardiol. 20, 483–499 (1992)
4. Karma, A.: Electrical alternans and spiral wave breakup in cardiac tissue. Chaos 4, 461–472 (1994)
5. Arce, H., Xu, A., González, H., Guevara, M.R.: Alternans and higher-oder rhythms in an ionic model of a sheet of ischemic ventricular muscle. Chaos 10, 411–426 (2000)
6. van Capelle, F.J., Durrer, D.: Computer simulation of arrhythmias in a network of coupled excitable elements. Circulation Research 47, 454–466 (1980)
7. Bub, G., Glass, L., Publicover, N.G., Shrier, A.: Bursting calcium rotors in cultured cardiac myocite monolayers. Proc. Natl. Acad. Sci. USA 95, 10283–10287 (1998)
8. Bub, G., Tateno, K., Shrier, A., Glass, L.: Spontaneous initiation and termination of complex rhythms in cardiac cel culture. J. Cardiovasc. Electrophysiol. 14, S229–S236 (2003)

Spiking Neural Network for On-line Cognitive Activity Classification Based on EEG Data

Stefan Schliebs, Elisa Capecci*, and Nikola Kasabov

KEDRI, Auckland University of Technology, New Zealand
{sschlieb,ecapecci,nkasabov}@aut.ac.nz

Abstract. The paper presents a method for the classification of EEG data recorded in two cognitive scenarios, a relaxing and memory task. The method uses a reservoir of spiking neurons that are activated by the spatio-temporal EEG data. The states of the reservoir are periodically read out and classified producing in a continuous classification result over time. After suitable optimization of the model parameters, we achieve a test accuracy of 82% on a small data set. Future applications of the proposed model are discussed including its use for an early detection of a cognitive impairment such as in Alzheimers disease.

Keywords: Spiking Neural Networks, Liquid State Machines, Reservoir Computing, EEG data classification, Cognitive tasks.

1 Introduction

Intellectual functioning including memory testing is a commonly used diagnosis tool to characterize the state of cognitive impairments such as Alzheimer's disease. In this paper, we investigate the idea to use the classification ability of a machine learning algorithm as an indicator for the detection of memory related cognitive diseases. We have collected EEG recordings from a single healthy subject performing a relaxing and a memory task; the latter represents the cognitive scenario. If the subject is healthy, a distinct difference between the EEG recordings of the two scenarios is expected and a classification algorithm should be able to tell the memory and relax scenarios reliably apart. Therefore, if a high classification accuracy is observed, the subject is expected to be healthy. On the other hand, if the classification performance is poor, it may be an indicator for memory related cognitive disease. In this paper, we investigate a brief proof of concept only. We are especially interested in establishing the suitability of a reservoir computing approach for the described learning scenario. Reservoir computing has reported promising results on the detection of epileptic seizures [1] and the classification of motor imagery based on EEG data streams [6]. While the above studies have investigated the suitability of Echo State Networks [4], we explore Liquid State Machines (LSM) [7] for classifying spatio-temporal EEG signals in this paper.

[*] Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 55–62, 2013.

2 SNN Model and Experimental Setup

An LSM consists of two main components, a “liquid” (also called reservoir) in the form of a recurrent Spiking Neural Network (SNN) [3] and a trainable readout function. The liquid is stimulated by spatio-temporal input signals causing neural activity in the SNN that is further propagated through the network due to its recurrent topology. Therefore, a snapshot of the neural activity in the reservoir contains information about the current and past inputs to the system.

The function of the liquid is to accumulate the temporal and spatial information of all input signals into a single high-dimensional intermediate state in order to enhance the separability between network inputs. The readout function is then trained to transform this intermediate state into a desired system output.

2.1 EEG Data Related to a Cognitive Memory Task

EEG data was collected using a standard 14-channel EEG recording device[1]. The tool is affordable, easy to transport and users do not need to be experts to manipulate it. The data was collected following two scenarios which were labelled as either the “relax” or the “memory” scenario. The EEG data was recorded from a single healthy subject over a period of 40 seconds for each of the two scenarios. The length of a session was chosen in accordance with the duration of the memory task. Brief test periods are preferred because they are more reliably to reproduce even if the participants are affected by a particular disorder [2]. The experiment labelled “Relax” was recorded with closed eyes, in order to avoid disturbing artefacts from blinking and the subject was asked to avoid thinking or planning thoughts as much as possible. For the “memory” experiment the Stenberg’s Memory Scanning Test (SMT) was adopted. The experiment was performed using the NBS Presentation software[2]. The SMT method is used in a wide range of scientific areas as it is an easy and practical model for evaluating information processing in working memory [2]. Both scenarios were each repeated for five times resulting in $2 \times 5 = 10$ sessions with a total of 400 seconds of recorded data altogether.

Fig. 1 depicts the recorded EEG time series for each channel and each session. Each plot contains two EEG traces, one for each class label (either “relax” or “memory” scenario). Due to the limitations of the used EEG device and the more or less informal recording conditions, some parts of the data might be impacted by artefacts and noise (cf. e.g. channel 8 in session 4). For the further processing, the recorded data was normalized by scaling it to zero mean and unit standard deviation and extreme outliers (values outside the first and 99-th percentile) were replaced by the mean of the time series of the corresponding channel.

The ten sessions were concatenating into a multiple time series containing the 14 EEG channels. The data mining task studied in this paper is to classify the

[1] Emotiv EPOC, `www.emotiv.com`

[2] version 0.7, `www.neurobs.com`

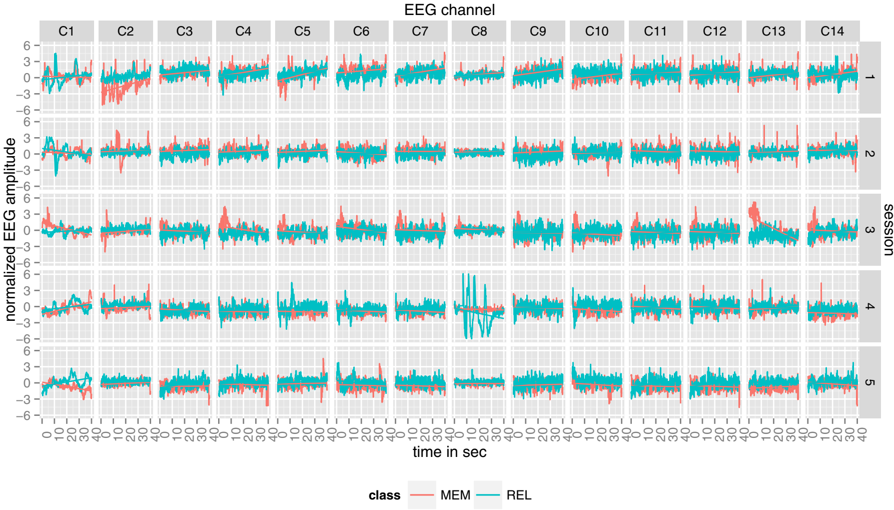


Fig. 1. EEG data recorded in 2 × 10 sessions (five sessions for each task) using a 14-channel (C1 to C14) EEG recording device

data on-line while it streams into a classification algorithm. Our goal is to be able to detect the scenario type of a (short) sequence of EEG activity using a Liquid State Machine approach.

2.2 Encoding EEG Data into Spike Sequences

The time series data obtained from the EEG device is presented to the reservoir in the form of an ordered sequence of real-valued data vectors. In order to obtain an input compatible with the SNN, each real value of a data vector is transformed into a spike train using a spike encoding. In [11], the authors explored two different encoding schemes, namely Ben's Spike Algorithm and a population encoding technique. Since only the latter one reported satisfying results on a temporal classification task, we restrict our analysis to this technique. Population encoding uses more than one input neuron to encode a single time series. The idea is to distribute a single input to multiple neurons, each of them being sensitive to a different range of real values. Our implementation is based on arrays of 50 receptive fields with overlapping sensitivity profiles as described in [9]. We refer to the mentioned references for further details and examples of this encoding algorithm.

As a result of the encoding, input neurons emit spikes at predefined times according to the presented data vectors. The input neurons are connected with a random set of reservoir neurons. The connection weights between input and reservoir neurons are initialized uniformly in the range $[-1, 1]$nA and then scaled using a linear scaling factor w_{in}. As a consequence, after scaling the input weights are in the range $[-w_{\text{in}}, w_{\text{in}}]$ nA. Configuring parameter w_{in} is very important, since it determines how strong the state of the reservoir neurons is influenced by the input signal. A low input scaling factor decreases the influence of the input signal and increases the influence of the recurrently connected reservoir neurons. Thus, by carefully adjusting parameter w_{in}, we decide how strongly the network responds to the input and how strongly it reacts to the activity generated by its reservoir neurons.

2.3 SNN Reservoir Structure

For the reservoir, we employ the Leaky Integrate-and-Fire neuron with exponential synaptic currents and a dynamic firing threshold [10] along with dynamic synapses based on the short-term plasticity (STP) proposed by Markram et al. [8]. We generate a recurrent SNN by aligning 1000 neurons in a three-dimensional grid of size $10 \times 10 \times 10$ in which the neurons are interconnected using the small-world pattern described in [7]. More details on the creation of this network and a complete description of all SNN parameters are given in [11].

2.4 Readout and Learning

A typical readout function convolves every spike by a kernel function which transforms the spike train of each reservoir neuron into a continuous analogue signal. We use an exponential kernel with a time constant of $\tau = 50$ ms. The convolved spike trains are then sampled using a time step of 10 ms resulting in 1000 time series – one for each neuron in the reservoir. In these series, the data points at time t represent the readout for the presented input sample. Readouts were labelled according to their readout time t. If the readout occurred at a time corresponding to either a relax or a memory session, then the corresponding readout is labelled accordingly.

The final step of the LSM framework consists of a mapping from a readout sample to a class label. The general approach is to employ a machine learning algorithm to learn the correct mapping using the readout data. Since the readout samples are expected to be linearly separable with regard to their class label [7], a comparably simple learning method can be applied for this task. From the labelled readouts, we compute a ridge regression model for mapping a reservoir readout sample to the corresponding class label. Ridge regression is essentially a regularized linear regression that can counteract model over-fitting by adjusting a regularization parameter α. We explored suitable configuration of this parameter in the experiments presented below.

3 Modelling EEG Data and Experimental Results

The LSM has a large number of parameters which have to be carefully adjusted in order to obtain satisfying classification results. We have worked with these type of neural networks in other studies [10,11] and we have identified some critical variables which require careful optimization. These variables are the scaling w_{in} of the input weights, the scaling w_{s} of the connection weights of the reservoir, the connection density λ of the reservoir neurons and the regularization parameter α that is used for learning the mapping of the reservoir state to the class label.

3.1 Parameter Optimization

We have performed a grid search in which 400 network configurations are evaluated regarding their test accuracy. We investigate all possible combinations of the following parameter options: $w_{\text{in}} \in [5, 10, 20, 40, 60]$, $w_{\text{s}} \in [5, 10, 20, 40, 80]$, $\lambda \in [3, 8]$ and $\alpha \in [0, 1, 5, 10, 20, 50, 75, 100]$.

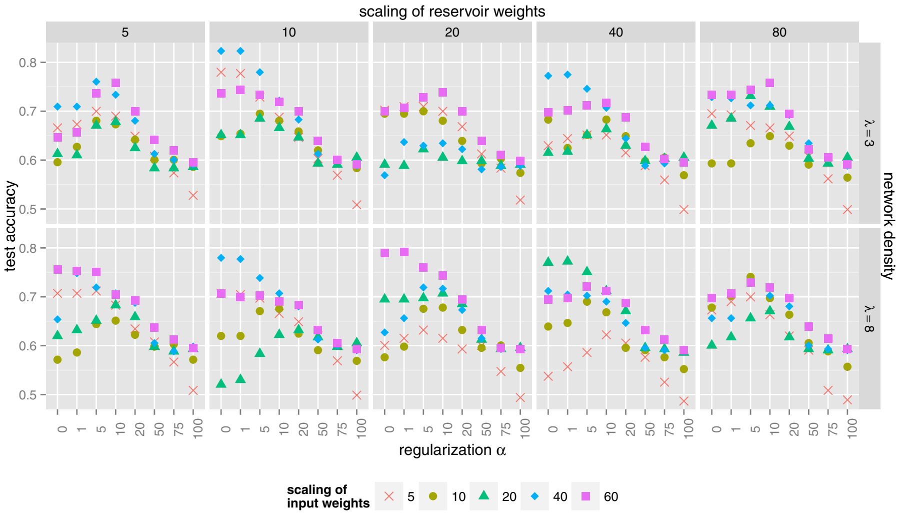


Fig. 2. Grid search for suitable parameter configurations for the LSM

The results of this grid search is reported in Fig. 2. Each point in this plot represent one of the 400 configuration tested during the grid search. The y-axis represents the test accuracy of the trained model which is our performance metric for this study. Since the data is perfectly balanced (identical number of instances for both classes), a test accuracy of 50% corresponds to a random classification of the data.

In the figure, we clearly note the impact of the regularization parameter α which directly controls the generalization capabilities of the trained model. If the regularization is too small/large the regression model over/under-fits the data. Also the input scaling has a considerate influence on the working of the model although the rule of this influence is less clear. The network density does not seem very important for this data set. From the grid search we select $\alpha = 1$, $w_{\text{in}} = 40$, $w_{\text{s}} = 10$ and $\lambda = 3$ as the most suitable configuration.

3.2 EEG Classification

Fig. 3 shows the outputs obtained from each of the individual processing steps of the LSM framework using the findings obtained in the previous section. Our data consists of a set of 14 time series over a time window of 2 × 5 × 40 seconds sampled at 128Hz. We have down-sampled this time series to 13Hz (keeping only every 10-th data sample) resulting in 5160 data frames. A frame was fed every 1ms as an input to the LSM which in turn was simulated for 5160ms, cf. Fig. 3A.

The encoded spike trains (population encoding) derived from the time series are depicted in 3B. The figures show a raster plot of the neural activity of the input neurons over time. A point in these plots indicates a spike fired by a particular neuron at a given time.

The obtained spike trains were then fed into a reservoir resulting in characteristic response patterns of the reservoir neurons, cf. Fig. 3C. The reservoir is continuously read out every 10ms of the simulation time using the technique described in section 2.4. Fig. 3D shows the readouts over time for the population-encoded reservoir inputs. The color in these plots indicates the value of the readout obtained from a certain neuron; the brighter the color, the larger the readout value.

The learning and classification step of the LSM framework is presented in the last plot of Fig. 3. The ridge regression model was trained on the first 3100 time points of readout data (60% of the entire time series) and then tested on the remainder of the time series. The raw output of the regression model was smoothed using a moving average with a small window size of 10 time points. Finally, we discretized the smoothed output using a simple threshold model with a cutoff value of 0.5. More specifically, if the regression reported a value larger than 0.5, then the sample was classified as the “memory class” and otherwise as the “relax class”. The model reported an expected excellent classification on the training data (100%) and a satisfying classification accuracy on the testing data (82.3%). The alternating pattern of the relax and the memory sessions is clearly visible from the model output and additional post-processing could potentially further improve the results. Despite possible over-fitting of the regression model, we could not improve the test accuracy by simply increasing the regularization parameter α.

4 Conclusions and Future Directions

From the here presented results, it seems principally possible to distinguish between the described two cognitive scenarios. Considering the short training period, the noisy nature of the data, its rather informal collection and the technical limitations of the EEG reading device, the initial results seem very acceptable. However, more extensive evidence is needed to establish the feasibility of a purely data driven diagnosis method for memory related diseases. Further improvement of the classification accuracy could be obtained by replacing the simple regression learning with more sophisticated learning techniques, e.g. the recently introduced NeuCube architecture [5], and through the automatic selection of relevant

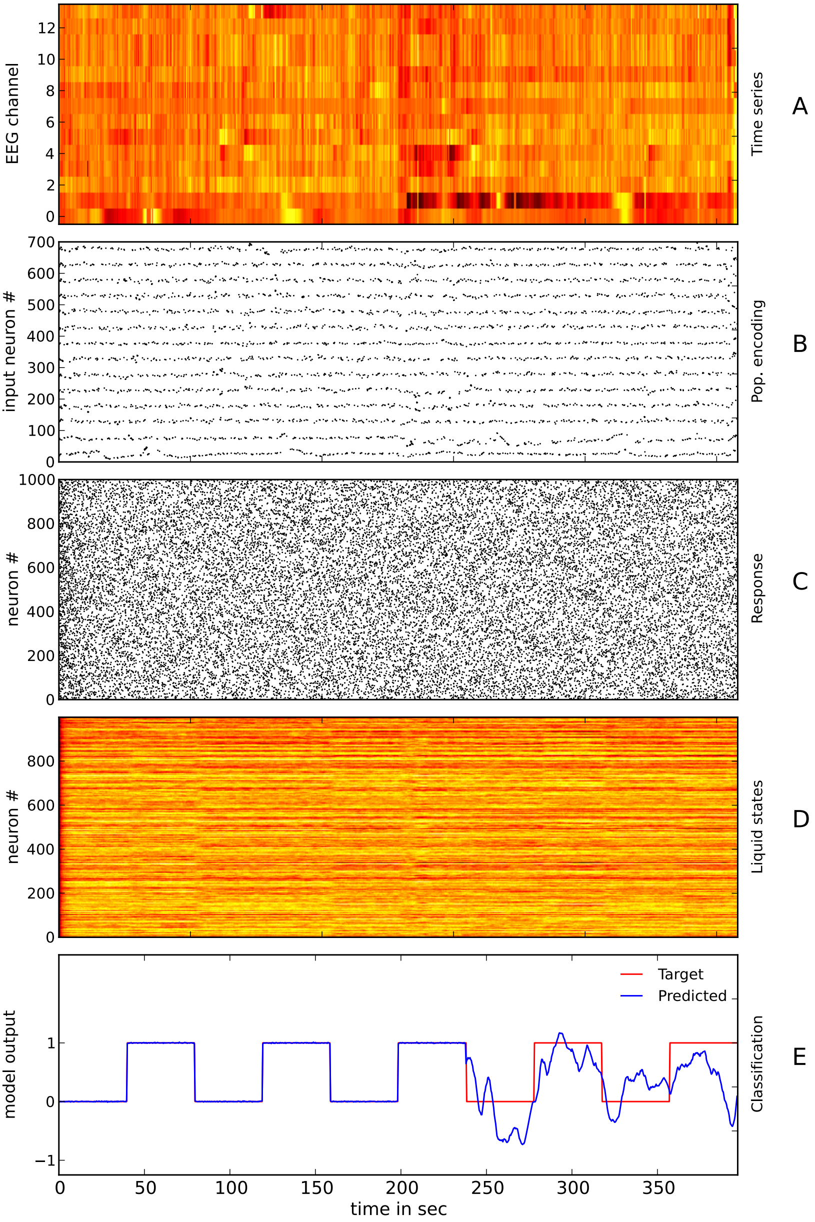


Fig. 3. Experimental results obtained from the activity recognition

features (EEG channels). Future work could involve the acquisition of a more suitable data set that also involves the EEG recordings from subjects suffering from cognitive diseases.

Acknowledgements. The research is supported by the Knowledge Engineering and Discovery Research Institute (www.kedri.aut.ac.nz) of the Auckland University of Technology.

References

1. Buteneers, P., Verstraeten, D., Nieuwenhuyse, B.V., Stroobandt, D., Raedt, R., Vonck, K., Boon, P., Schrauwen, B.: Real-time detection of epileptic seizures in animal models using reservoir computing. Epilepsy Research 103(2-3), 124–134 (2013)
2. von der Elst, W., van Boxtel, M.J., van Breukelen, G.P., Jolles, J.: Assessment of information processing in working memory in applied settings: the paper and pencil memory scanning test. Psychological Medicine 37, 1335–1344 (2007)
3. Gerstner, W., Kistler, W.M.: Spiking Neuron Models: Single Neurons, Populations, Plasticity. Cambridge University Press, Cambridge (2002)
4. Jaeger, H.: The "echo state" approach to analysing and training recurrent neural networks. Tech. rep., Fraunhofer Institute for Autonomous Intelligent Syst. (2001)
5. Kasabov, N.: Neucube evospike architecture for spatio-temporal modelling and pattern recognition of brain signals. In: Mana, N., Schwenker, F., Trentin, E. (eds.) ANNPR 2012. LNCS (LNAI), vol. 7477, pp. 225–243. Springer, Heidelberg (2012)
6. Kindermans, P.J., Buteneers, P., Verstraeten, D., Schrauwen, B.: An uncued brain-computer interface using reservoir computing. In: Workshop: Machine Learning for Assistive Technologies, Proceedings, p. 8. Ghent University, Department of Electronics and information systems (2010)
7. Maass, W., Natschläger, T., Markram, H.: Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation 14(11), 2531–2560 (2002)
8. Markram, H., Wang, Y., Tsodyks, M.: Differential signaling via the same axon of neocortical pyramidal neurons. Proceedings of the National Academy of Sciences 95(9), 5323–5328 (1998)
9. Schliebs, S., Defoin-Platel, M., Kasabov, N.: Integrated feature and parameter optimization for an evolving spiking neural network. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008, Part I. LNCS, vol. 5506, pp. 1229–1236. Springer, Heidelberg (2009)
10. Schliebs, S., Fiasché, M., Kasabov, N.: Constructing robust liquid state machines to process highly variable data streams. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) ICANN 2012, Part I. LNCS, vol. 7552, pp. 604–611. Springer, Heidelberg (2012)
11. Schliebs, S., Hunt, D.: Continuous classification of spatio-temporal data streams using liquid state machines. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012, Part IV. LNCS, vol. 7666, pp. 626–633. Springer, Heidelberg (2012)

Spatio-temporal EEG Data Classification in the NeuCube 3D SNN Environment: Methodology and Examples

Nikola Kasabov[1,*], Jin Hu[2], Yixiong Chen[2], Nathan Scott[1], and Yulia Turkova[1]

[1] Knowledge Engineering and Discovery Research Institute,
Auckland University of Technology, New Zealand
nkasabov@aut.ac.nz

[2] State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract. A vast amount of complex spatio-temporal brain data, such as EEG-, have been accumulated. Technological advances in many disciplines rely on the proper analysis, understanding and utilisation of these data. In order to address this great challenge, the paper utilizes the recently introduced by one of the authors 3D spiking neural network environment called NeuCube for spatio-temporal EEG data classification. A methodology is proposed and illustrated on two small-scale examples: classifying EEG data for music- versus noise perception, and person identification based on music perception. Future development and usage of the NeuCube environment can be expected to significantly further the creation of novel brain-computer interfaces, cognitive robotics and medical engineering devices.

Keywords: EEG, spatio-temporal data, spiking neural networks, music perception, NeuCube.

1 Introduction

Over the last couple of decades a vast amount of information about structural and functional characteristics of the human brain has been accumulated [1,2,3,4,5]. An enormous quantity of Spatio-Temporal Brain Data (STBD) has been collected, such as: Electroencephalogram (EEG) [6,7], Magneto Encephalograph (MEG) [8], functional Magnetic Resonance Imagining (fMRI) [9,10,11], gene expression data related to brain states [12], *etc.* However, the analysis of this type of data presents a challenge to researchers. Traditional methods, such as Multiple Linear- and Logistic Regression, Support Vector Machines (SVM), Multilayer Perceptron Neural Networks, Hidden Markov Models, rule-based systems, *etc.* have been used with limited success [13] for the classification of EEG data [6,7]. All these methods emphasise either the spatial or temporal component of the data, but do not take into account their dynamic interaction, and

[*] Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 63–69, 2013.

neither can they accommodate multimodal STBD and prior information about the source of this data. They are therefore less appropriate for the classification and interpretation of brain data than the approaches discussed below.

The brain represents and processes information in the form of many trains of temporal electrical potentials that can be considered binary events (spikes) and are transferred between neurons through synaptic connections. Through learning from data the synaptic connections are modified to reflect more precisely the timing of the data from the sensory inputs. And this is one of the principles of spiking neural networks (SNN), considered the third generation of brain-inspired neural network techniques [16]. SNN methods and engineering systems have been developed for: learning from data [14,15,16,17]; system design and implementation [18,19]; encoding continuous input data into spike trains, such as the silicon retina [20] and the silicon cochlea [21] sensory devices; neurogenetic computation [22,23]; high performance and neuromorphic engineering systems and supercomputers [24,25].

Promising features of SNN are: compact representation of space and time; fast information processing; time-based and frequency-based information representation. They are therefore more appropriate for use with brain data, as this is inherently spatio- and spectro-temporal. Recently novel SNN methods for spatio-temporal pattern recognition were developed [31]. Among them are two types of evolving SNN classifiers (deSNN [26] and SPAN [27]), and pilot applications for moving object recognition and simple EEG data classification[1]. This paper develops further this research towards EEG STBD classification using the framework recently proposed by Kasabov for STBD – NeuCube [28].

2 The NeuCube Framework for STBD

The NeuCube framework (fig.1) consists of:

- A module for encoding input data into spike sequences. Input STBD are encoded into trains of spikes that capture the data temporal characteristics using for example some of the following methods: Population Coding [19]; Address Event Representation (AER) [20,21]; Bens Spike Algorithm [26]. In the AER method for example a spike (either positive or negative) is generated if only the difference between two consecutive values of an input variable is above a defined threshold, thus capturing temporal differences in the data.
- A 3D SNN reservoir (SNNr) of leaky integrate and fire model (LIFM) neurons. The spike trains are entered into the reservoir (SNNr) where each neuron has predefined 3D spatial coordinates. The STDP learning rule [14] is applied that will allow the SNNr to create connections based on temporal associations between input spikes.
- Output classification module. Neurons from the 3D SNNr are connected to neurons in the output classification module and spiking activity patterns of the SNNr are continuously passed as input information to the classifier.

[1] `http://ncs.ethz.ch/projects/evospike`

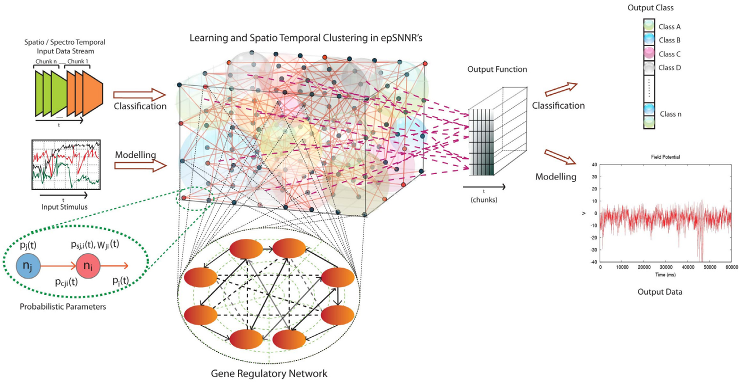


Fig. 1. A block diagram of the NeuCube framework (from [28])

- Gene Regulatory Network (GRN) module for parameter control and optimisation of SNNr (optional). GRN models can also be added as additional parameters to control the activity of the neurons in the SNNr.

NeuCube utilises for the first time the following principles for the design and development of a system for STBD modelling and pattern recognition:

1. The same paradigm, spiking information processing, that generates STBD at a low level of brain information processing, is used to represent and to process data.
2. An information model, created for STBD, will not only use the available data, but also using prior knowledge, namely structural and functional information about the source of the data. The SNN system has a spatial structure that approximates spatially located areas of the brain where STBD is collected.
3. Brain-like learning rules, such as STDP [14], are used to learn temporal cause-effect relationships between spatially distributed neurons in the SNN system to reflect on spatio-temporal interaction in the input data.
4. The system is evolving as previously unknown classes could be added incrementally as a result of new STBD patterns being learned and recognised, which is also a principle of brain cognitive development [29].
5. The system will always retain a transparent spatio-temporal memory that can be mined and interpreted either in real time or retrospectively for new knowledge discovery.
6. The system is able to recognise and predict the outcome of a new STBD pattern that is similar to previously learnt ones even before the new pattern is fully presented to its inputs.

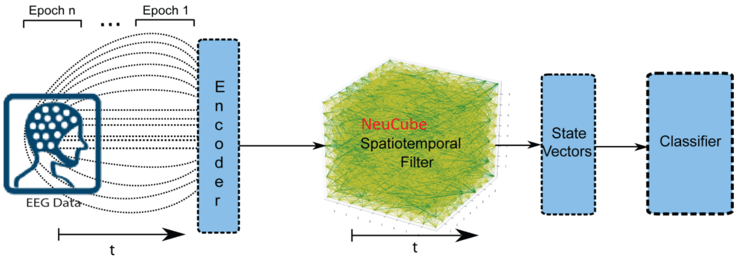
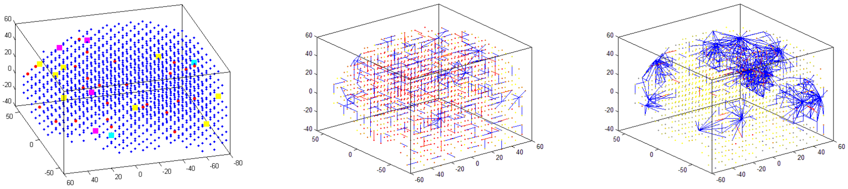


Fig. 2. A block diagram of using NeuCube for EEG STBD classification (from [28])



(a) Spiking activity

(b) Connectivity before training – small world connections

(c) Connectivity after training

Fig. 3. Illustrative visualisation of connectivity and spiking activity of a SNNr that consists of 1471 neurons

3 Methodology for EEG STD Classification in the NeuCube Environment

The following methodology is proposed here for EEG STD classification implementing the block diagram from Fig. 2 with the use of NeuCube:

1. Collected EEG signals are encoded into trains of spikes using the Address Event Representation (AER) method.
2. These spike trains are entered into a reservoir SNNr with 1471 spiking neurons. For EEG we use the mapping of EEG channels into the Talairach template coordinates from [30]. The 3D coordinates of the neurons in the SNNr correspond to the Talairach coordinates [2] so that any EEG data for an arbitrary human subject can be mapped into the 3D SNNr.
3. All 1471 neurons from the 3D SNNr are connected to neurons in the output classification module and spiking activity patterns of the SNNr are continuously passed as input information to the classifier. As a classifier we use deSNN [26].
4. EEG STBD is learned in the 3D SNNr in an unsupervised mode using the STDP learning rule [14].

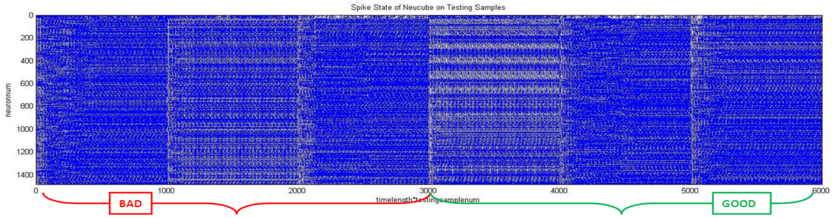


Fig. 4. Example spiking activity over time for the example experiment

5. After that the training EEG STBD is propagated again and a classifier is trained in a supervised mode to recognise labelled patterns from the SNNr.
6. The system is tested on test EEG STBD.
7. If necessary, parameters are optimised for a better performance
8. The system is analysed for a better EEG STBD understanding and new knowledge discovery.

As an illustration, Fig. 3 shows the spiking activity (a) and connectivity of a SNNr of 1471 neurons trained on EEG STBD collected through a 14 channel EEG wireless headset (Emotiv) before training - (b) and after training - (c). It can be seen that as a result of training, new connections have been created that represent spatio-temporal interaction between EEG channels captured from the data. The connectivity can be viewed dynamically for every new pattern submitted.

4 Examples of EEG STBD Classification in NeuCube

As an illustration of the above methodology here we use two examples of EEG data classification of music perception. Data was collected with the use of the 14 Emotiv channels: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4, as labelled in the standard International 10-20 scheme. Two scenarios are presented.

In the first scenario EEG data was collected from a single subject with the objective to classify EEG data of music versus noise perception. Each of the stimuli was presented 6 times for a duration of 10 seconds each, altogether 12 sessions. Six sessions were used for training the NeuCube system and six for testing (Fig. 4). The test results indicated 100% accurate classification on training data and 83.33% classification accuracy on test data (one session of music was classified incorrectly as noise).

In a second scenario, the research question was if perception of music can be used to identify a person. EEG data from two subjects was collected in the same way above, with the same music used as stimulus. The trained NeuCube can identify the subjects based on musical response with 100% accuracy on training data and 83.33% accuracy on test data (only one miss-classified session). When the system was trained on the EEG perception of the two subjects data of the noise stimulus, the system could not discriminate the subjects. The experiments above are of a very small scale and any hypotheses drawn from these experiments need to be further tested in the future.

5 Conclusion

The paper presents a methodology and illustrative examples for EEG STBD classification with the use of the NeuCube framework [28]. The conclusion is that the NeuCube has the potential to become a valuable environment for classification of EEG data and for a better understanding of the data. More experiments will be conducted in the future for a detailed comparative analysis, for a larger scale of EEG data, and for investigation of more complex cognitive problems. The application of the method for the design of new BCI will also be studied.

Acknowledgements. This research is funded by the Ministry of Business, Innovation and Enterprise (MBIE) of New Zealand through a Strategic Alliance New Zealand-China grant, and also by the Chinese Academy of Sciences

References

1. Zillies, K., Amunts, K.: Centenary of Brodmann's map – conception and fate. Nature Reviews Neuroscience 11, 139–145 (2010)
2. Talairach, J., Tournoux, P.: Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging. Thieme Medical Publishers, NY (1988)
3. Evans, A.C., Collins, D.L., Mills, S.R., et al.: 3D statistical neuroanatomical models from 305 MRI volumes. In: Proc. IEEE-Nuclear Science Symp. Medical Imaging Conference, pp. 1813–1817 (1993)
4. Toga, A., Thompson, P., Mori, S., et al.: Towards multimodal atlases of the human brain. Nature Reviews Neuroscience 7, 952–966 (2006)
5. Abeles, M.: Corticonics. Cambridge University Press, NY (1991)
6. Fiasché, M., Schliebs, S., Nobili, L.: Integrating Neural Networks and Chaotic Measurements for Modelling Epileptic Brain. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) ICANN 2012, Part I. LNCS, vol. 7552, pp. 653–660. Springer, Heidelberg (2012)
7. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., et al.: A review of classification algorithms for EEG-based brain-computer interfaces. J. Neural Eng. 4(2), 1–15 (2007)
8. Stam, C.J.: Functional connectivity patterns of human magnetoencephalographic recordings: A small-world network? Neurosci. Lett. 355, 25–28 (2004)
9. De Charms, R.C.: Applications of real-time fMRI. Nature Reviews Neuroscience 9, 720–729 (2008)
10. Mitchel, T., Hutchinson, R., et al.: Learning to Decode Cognitive States from Brain Images. Machine Learning 57, 145–175 (2004)
11. Broderson, K., Wiech, K., Lomakina, E., et al.: Decoding the perception of pain from fMRI using multivariate pattern analysis. NeuroImage 63, 1162–1170 (2012)

12. Hawrylycz, M., et al.: An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489, 391–399 (2012)
13. Gerstner, W., Sprekeler, H., Deco, G.: Theory and simulation in neuroscience. Science 338, 60–65 (2012)
14. Song, S., Miller, K., Abbott, L., et al.: Competitive hebbian learning through spike-timing-dependent synaptic plasticity. Nature Neuroscience 3, 919–926 (2000)
15. Thorpe, S., Gautrais, J.: Rank order coding. Comput. Neuroscience: Trends in Research 13, 113–119 (1998)
16. Maass, W., Natschlaeger, T., Markram, H.: Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural Computation 14(11), 2531–2560 (2002)
17. Izhikevich, E.: Polychronization: Computation with Spikes. Neural Computation 18, 245–282 (2006)
18. Belatreche, A., Maguire, L.P., McGinnity, M.: Advances in Design and Application of Spiking Neural Networks. Soft Comput. 11(3), 239–248 (2006)
19. Gerstner, W.: What's different with spiking neurons? In: Mastebroek, H., Vos, H. (eds.) Plausible Neural Networks for Biological Modelling, pp. 23–48. Kluwer Academic Publishers (2001)
20. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128x128 120dB 30mW Asynchronous Vision Sensor that Responds to Relative Intensity Changes. ISSCC Digest of Technical Papers, pp. 508–509 (2006)
21. Liu, S.C., Delbruck, T.: Neuromorphic sensory systems. Curr. Opinion in Neurobiology 20(3), 288–295 (2010)
22. Benuskova, L., Kasabov, N.: Computational neuro-genetic modelling. Springer, New York (2007)
23. Kasabov, N.: To spike or not to spike: A probabilistic spiking neuron model. Neur. Netw. 23(1), 16–19 (2010)
24. Furber, S.: To Build a Brain. IEEE Spectrum 49(8), 39–41 (2012)
25. Indiveri, G., Horiuchi, T.K.: Frontiers in neuromorphic engineering. Frontiers in Neuroscience 5, 1–2 (2011)
26. Kasabov, N., Dhoble, K., Nuntalid, N., Indiveri, G.: Dynamic Evolving Spiking Neural Networks for On-line Spatio- and Spectro-Temporal Pattern Recognition. Neural Networks 41, 188–201 (2013)
27. Mohemmed, A., Schliebs, S., Matsuda, S., Kasabov, N.: SPAN: Spike Pattern Association Neuron for Learning Spatio-Temporal Sequences. Int. J. of Neural Systems 22(4), 1–16 (2012)
28. Kasabov, N.: NeuCube EvoSpike Architecture for Spatio-Temporal Modelling and Pattern Recognition of Brain Signals. In: Mana, N., Schwenker, F., Trentin, E. (eds.) ANNPR 2012. LNCS (LNAI), vol. 7477, pp. 225–243. Springer, Heidelberg (2012)
29. Kasabov, N.: Evolving connectionist systems: The knowledge engineering approach. Springer (2007)
30. Koessler, L., Maillard, L., Benhadid, A., et al.: Automated cortical projection of EEG sensors: Anatomical correlation via the international 10–10 system. NeuroImage 46, 64–72 (2006)
31. Kasabov, N.: Evolving Spiking Neural Networks and Neurogenetic Systems for Spatio- and Spectro-Temporal Data Modelling and Pattern Recognition. In: Liu, J., Alippi, C., Bouchon-Meunier, B., Greenwood, G.W., Abbass, H.A. (eds.) WCCI 2012. LNCS, vol. 7311, pp. 234–260. Springer, Heidelberg (2012)

NeuCubeRehab: A Pilot Study for EEG Classification in Rehabilitation Practice Based on Spiking Neural Networks

Yixiong Chen[1], Jin Hu[1], Nikola Kasabov[2], Zengguang Hou[1], and Long Cheng[1]

[1] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing
`{yixiong.chen,jin.hu,zengguang.hou,long.chen}@ia.ac.cn`

[2] Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland
`nkasabov@aut.ac.nz`

Abstract. One of the most important issues among active rehabilitation technique is how to extract the voluntary intention of patient through bio-signals, especially EEG signal. This pilot study investigates the feasibility of utilizing a 3D spiking neural networks-based architecture named NeuCube for EEG data classification in the rehabilitation practice. In this paper, the architecture of the NeuCube is designed and a Functional Electrical Stimulation (FES) rehabilitation scenario is introduced which requires accurate classification of EEG signal to achieve active FES control. Three classes of EEG signals corresponding to three imaginary wrist motions are collected and classified. The NeuCube architecture provides promising classification results, which demonstrates our proposed method is capable of extracting the voluntary intention in the rehabilitation practice.

Keywords: Spiking Neural Network, Rehabilitation, EEG classification, FES, NeuCube.

1 Introduction

Patients who have Spinal Cord Injury (SCI) or stroke always suffer from motor disorder, poor blood circulation on the affected limbs as well as psychological problems, since their ability of controlling the limbs is weakened caused by muscle denervation. Rehabilitation exercises are required to restore the lost movement function of paralyzed limbs, which help patients out of the predicament physiologically and psychologically [1]. Traditional rehabilitation methods are executed by moving the paralyzed limbs repeatedly with the help of the physical therapist, which is time-consuming and laborious [2]. Due to the rapid development of modern technology, this manual treatment is gradually replaced by rehabilitation equipments such as rehabilitation robot and Functional Electrical Stimulation (FES) device, etc. These devices are designed for repetitive movement exercises for paraplegic or hemiplegic patients. These advanced rehabilitation techniques have enhanced the effect of rehabilitation excises, especially when the patient's intention of how to move the limb and the actual movement (either

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 70–77, 2013.

assisted by rehabilitation robot or induced by FES) is coordinated. This is the concept of active rehabilitation introduced recent years [3].

To achieve this active rehabilitation strategy, the critical issue lies on the extraction of the patient's voluntary intention. Traditional methods utilize force/torque sensors to measure the active force/torque of patients applied on the rehabilitation devices. However, SCI or stroke patients don't have sufficient muscle strength in most cases. On the other hand, bio-signals, such as surface Electromyography (sEMG) and Electroencephalography (EEG) have been utilized as bio-feedback to improve the efficiency of rehabilitation excises and can be also used to extract intention of the patients [4, 5]. Among all kinds of bio-signals, EEG signal is the most direct indicator of brain activities, and its signal strength doesn't decline when it comes to SCI or stroke patients, making EEG the ideal bio-signal for voluntary intention extraction.

Variety of algorithms using EEG data for intention extraction have been developed recently, such as Support Vector Machines (SVM), traditional artificial neural networks, hidden Markov models, etc. But they have limited capacity to achieve the integration of complex and long temporal spatial/spectral components because they usually either ignore the temporal dimension or over-simplify its representation [6]. However, Spiking Neural Networks (SNN) which is a dynamic system, using trains of spikes to encode the data, has the potential to perform computation on temporal patterns [7]. In this paper, we implement an SNN named NeuCube [6,8,9] proposed by one of the listed authors to classify three classes of EEG data in an FES-involved rehabilitation scenario, and this research is a pilot study for future design of active rehabilitation strategy. The rest of this paper is organized as follows. In Section 2, the scenario of this pilot study as well as the architecture of NeuCube are described in detail. Then classification results are presented in Section 3. Finally, conclusion and future work are drawn in Section 4.

2 Methods

2.1 Scenario Description and Data Collection

This pilot study concerns the application of NeuCube architecture to classify EEG data which is a Spatio-Temporal Brain Data (STBD), and the classification results are utilized as the control signal for FES therapy. FES is a commonly used therapy method for stroke and SCI patients. FES involves artificially inducing a current in the specific motor neurons to generate muscle contractions which has shown satisfactory results in movement restoration and neuro rehabilitation [11]. In traditional FES therapy, current stimulus with predefined intensity is applied on the limbs of paralyzed patients to help them achieve certain limb movement. Nevertheless, without regarding voluntary intention of patients, the rehabilitation effect is limited in traditional FES therapy, for the reason that this purely passive process is less likely to inspire patients to be involved in the training. To improve rehabilitation effect, human-in-loop FES control is introduced, in which the patients' intention is extracted and then used to control the FES intensity to achieve certain induced contraction of muscles according to the willing of the patients.

Figure 1 shows a case of human-in-loop FES control. The paralyzed patient lost his control of wrist, and FES is applied on the wrist extensor and flexor to assist patients in restoring the hand movement function. During the training, the intention of either extent wrist or flex wrist is classified by NeuCube architecture using the EEG data of the patient. The classification result is employed to adjust FES parameters to certain proper intensity so that the desired wrist position can be achieved. Fulfillment of this human-in-loop FES control entirely depends on the precise classification of the EEG data. Therefore, for a pilot study, we first focus on the classification phase and then take control phase into account. To demonstrate the capability of the proposed methodology, EEG data of three classes, i.e. EW (imagine extent wrist), FW (imagine flex wrist) and RP (imagine maintain wrist in natural and relaxed position) were collected and utilized in training as well as validation.

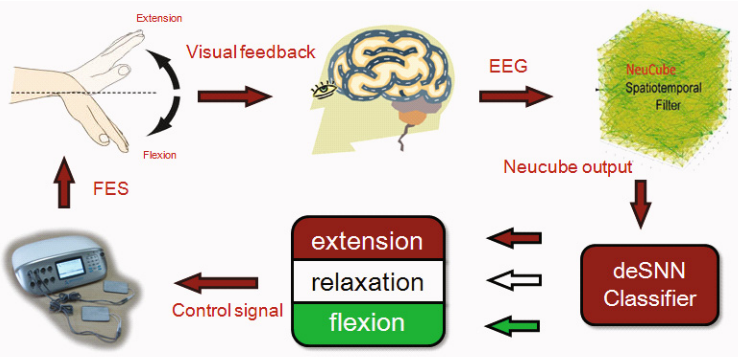


Fig. 1. A block of human in loop FES control based on NeuCube architecture

EEG data were collected by Emotive, which is a commercial EEG acquisition helmet with 14 channels. A healthy male subject was asked to repeatedly imagine three motions, i.e. EW, FW and RP after a sound ‘beep’ as a reminder with his eyes closed to minimize noise signals. Each class of imagination was repeated for 50 times, 1 sec for each time. Every sample contained 128 data points for each channels. The 150 samples were randomly and evenly divided into training group and validation group.

2.2 NeuCube Architecture

The NeuCube Architecture consists of three main modules: a brain-like SNN reservoir, an Address Event Representation (AER) encoding module and an output classification module [8]. The overall architecture is shown in Figure 2. The EEG data are collected and transferred into trains of spikes, which are passed into the SNN reservoir. According to Spike-Timing-Dependent-Plasticity (STDP) learning rule, the patten of the spatio-temporal EEG data can be remembered in the form of the spiking activities of all the neurons. The states of all the neurons in the reservoir are recorded and latter used to train the classifier. In this pilot study, we chose Dynamic Evolving Spiking Neural Networks (deSNN) as classifier [9].

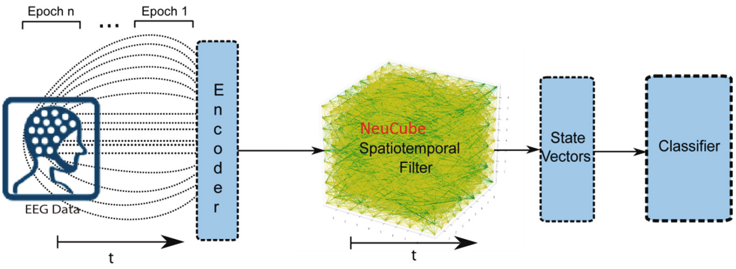
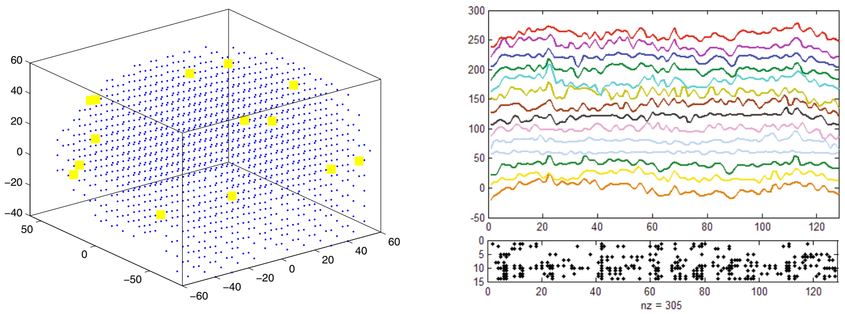


Fig. 2. Overall architecture of NeuCube for EEG data classification (from [6])

Reservoir Structure. The brain-like reservoir is a 3D SNN of 1471 neurons based on the Leaky Integrate and Fire Model (LIFM) [12], and it is a 3D approximate map of brain areas. 14 of these 1471 neurons are specified as the input neurons, from which the trains of spikes transferred from EEG data are propagated into the reservoir. The coordinates of these input neurons are as the same as the locations where the 14 channels of EEG data are collected [8]. Figure 3(a) is the visualization of the 3D structure of the reservoir. Those blue dots represent internal neurons, while those yellow ones represent 14 input neurons. The internal neurons of the reservoir utilize LIFM to calculated their states, i.e. spike or not spike. To initialize the connection weights between neurons in the reservoir, the Small World Connection (SWC) rule is applied. The spatial distance of two neurons is calculated to determine their initial connection weight. According to this rule, neurons within small area are more densely connected, and the weight of the connections are depended on the distance between the neurons [8]. Arbitrarily, 80 percent of connections are initialized to be exhibitory and the other 20 to be inhibitory.



(a) 3D structure of SNN reservoir

(b) EEG data of one sample collected by Emotive and trains of spikes encoded from these data using AER

Fig. 3. Reservoir structure and data encoding

EEG Data Encoding. To transfer the EEG data which are analogue values into trains of spikes, Address Event Representation (AER) algorithm is applied [13]. The difference of adjacent data points of the same input variable is calculated, once the difference reach certain threshold, a spike will be emitted at the input neuron which represents this particular channel. This algorithm is suitable for fast encoding compared with other encoding algorithm such as Bens Spike Algorithm (BSA) encoding or population encoding [14, 15]. Figure 3(b) shows EEG data of one sample collected by Emotive and trains of spikes encoded from these data using AER.

Unsupervised Learning Rule of the Reservoir. The unsupervised STDP learning rule, used here to train the brain-like reservoir, utilizes Hebbian plasticity. Whether synapses strengthened or weakened is based on the spike timing of pre-synaptic and post-synaptic neurons [10]. If pre-synaptic neuron spikes first, then the connection weight between the two neurons increases, otherwise it decreases. The smaller the time interval of those two spikes emitted by the pre- and post-synaptic neuron, the greater the modification of the connection weight is. The amount of synaptic modification $F(\Delta t)$ can be calculated as follows:

$$F(\Delta t) = \begin{cases} \mathrm{A}_{+} \exp(\Delta t/\tau) & \text{if } \Delta t < 0 \\ -\mathrm{A}_{-} \exp(\Delta t/\tau) & \text{if } \Delta t \geq 0 \end{cases} \tag{1}$$

where Δt is the time interval between the spike of pre- and post-synaptic neurons. The parameter τ determines the ranges over which the STDP become effective. A_{+} and A_{-} determine the maximum amounts of synaptic modification [10].

Output Classification Module. deSNN, which is employed as the output classification module of the NeuCube architecture, combines Rank Order (RO) and Spike Driven Synaptic Plasticity (SDSP) learning rules to evolve a new spiking neuron and new connections to learn new patterns from incoming data [8, 9]. For each training sample, a new output neuron is evolved, and this output neuron is labeled as the particularly class belonged to the training sample. To calculate the connection weight between this new evolved output neuron and all the deSNN neurons, both RO learning and SDSP learning rules are adopted. The RO learning sets the initial values of the connection weights, the SDSP rule dynamically adjusts these connection weights based on further incoming spikes. For the *i*th output neuron, at time *t*, the connection weight between this output neuron and the *j*th neuron of the deSNN can be calculated as follows:

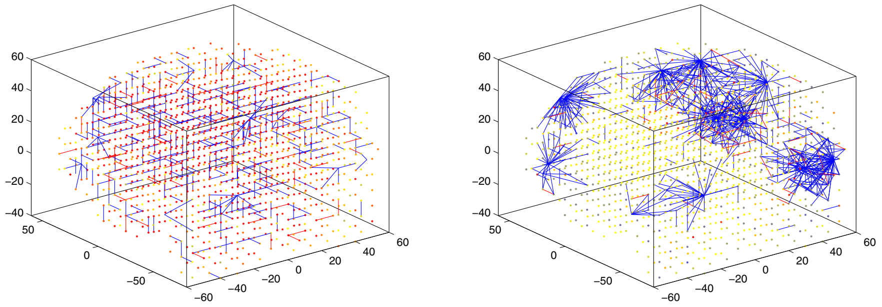
$$w_{j,i}(t) = \alpha \cdot mod^{order_{j,i}} + \sum_{k=1}^{t} e_j(k) \cdot D \tag{2}$$

where α is a factor which determines how much of the RO learning is involved in the calculation of the connection weight. *mod* is modulation factor and $order_{j,i}$ is the order of first spike from the *j*th neuron of the deSNN. $e_j = 1$ if there is a consecutive spike at synapse j at time k and $e_j = -1$ otherwise.

Learning of deSNN is one-pass, and the amount of output neuron equals to the amount of training samples. For a validation sample, another output neuron is created using the same method when training process is executed. Then the Euclidean distances between the weights vector of this new neuron and all the existed neurons are calculated, The new validation sample is associated with the closest output neuron based on the minimum distance between the weight vectors.

3 Results

150 samples which belonged to three different classes i.e. EW, FW and RP were randomly and evenly divided into training group and validation group. After training, the spatio-temporal patten has been captured from the EEG data which can be visualized by the connectivity change of the reservoir. Figure 4(a) shows the initial connections of the NeuCube reservoir before training process was executed. Those dots represent 1471 neurons, and those lines represent the major connections whose weights are greater than a certain threshold with the colors indicate either these connections are exhibitory or inhibitory (blue for exhibitory and red for inhibitory). Figure 4(b) shows the connections after the training process. Comparison of Figure 4(a) and Figure 4(b) reveals that the random connections before training became more regular and the connection densities of certain areas are strengthened especially where the input neurons locate and where the activity level is relatively high.



(a) Initial connections

(b) Connections after training

Fig. 4. Connection modification according to STDP learning rule

Figure 5 shows the output of 1471 neurons in NeuCube reservoir for the entire training process when all the training data were propagated. The output of NeuCube reservoir is used to train the deSNN classifier. Training and validation procedure were carried on for 20 times. The average classification accuracies using NeuCube architecture for EW, RP and FW are 88%, 83% and 71% respectively. Two comparison classification algorithms were also carried out. For the first comparison algorithm, only deSNN classifier was applied, and for the second one, the STDP learning mechanism for NeuCube reservoir was disabled. Experiment results which are presented in Figure 6 have shown the average accuracies of using these two comparison algorithms are less than using NeuCube architecture.

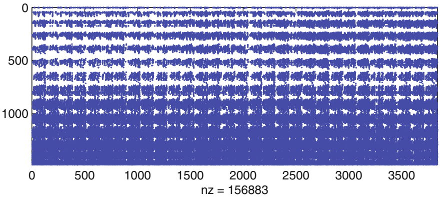


Fig. 5. Output of 1471 neurons in NeuCube reservoir for all the training data

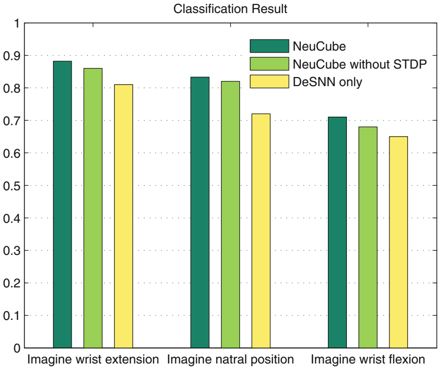


Fig. 6. Classification results

4 Conclusion and Future Work

In this paper, the NeuCube architecture is described and illustrated with a case study of EEG data classification for extracting the voluntary intention of the patient in FES rehabilitation practice. The NeuCube architecture provides promising classification results even with the data collected just from a 14-channels EEG acquisition equipment. More satisfactory classification results can be expected when the EEG data are collected from more precise EEG acquisition equipment with more channels. The potential of the NeuCube architecture to become an useful algorithm for the EEG data classification in rehabilitation practice has been proven.

In the future, focus will be on the optimization of the parameters to increase classification accuracy as well as to improve computing efficiency which is critical for real-time rehabilitation application. A larger scale of EEG data under more complex rehabilitation tasks will be collected using 64 channel EEG acquisition equipment to further validate the effectiveness of the NeuCube architecture. After that, the NeuCube architecture will be embedded into FPGA chip, so that the time cost of computing can satisfy the real-time and close-loop control for rehabilitation practice.

Acknowledgments. This research is supported in part by the International Cooperation Project of China (Grant 2011DFG13390).

References

1. Chen, Y., Hu, J., Zhang, F., Hou, Z.: Simulation Study of an FES-Involved Control Strategy for Lower Limb Rehabilitation Robot. In: Su, C.-Y., Rakheja, S., Liu, H. (eds.) ICIRA 2012, Part II. LNCS, vol. 7507, pp. 85–95. Springer, Heidelberg (2012)
2. Koji, I., Takahiro, S., Toshiyuki, K.: Lower-limb Joint Torque and Position Controls by Functional Electrical Stimulation. In: Complex Medical Engineering, pp. 240–249 (2006)

3. Li, J.T., Zheng, R.Y., Zhang, Y.R., Yao, J.C.: iHandRehab: an Interactive Hand Exoskeleton for Active and Passive Rehabilitation. In: 2011 IEEE International Conference on Rehabilitation Robotics, Zurich, pp. 1–6 (2011)
4. Oonishi, Y., Sehoon, O.: A New Control Method for Power-Assisted Wheelchair Based on the Surface Myoelectric Signal. IEEE Transaction on Industral Electronics 57(9), 3191–3196 (2010)
5. Ang, K.K., Guan, C., Phua, K.S.: Transcranial direct current stimulation and EEG-based motor imagery BCI for upper limb stroke rehabilitation. In: 34th Annual International Conference of the IEEE EMBS, San Diego, pp. 4128–4131 (2012)
6. Kasabov, N.: Evolving Spiking Neural Networks and Neurogenetic Systems for Spatio- and Spectro-Temporal Data Modelling and Pattern Recognition. In: 2012 IEEE World Congress on Computational Intelligence, pp. 234–260 (2012)
7. Buonomano, D., Maass, W.: State-dependent computations: Spatio-temporal processing in cortical networks. Nature Reviews, Neuroscience 10, 113–125 (2009)
8. Kasabov, N.: NeuCube EvoSpike Architecture for Spatio-Temporal Modelling and Pattern Recognition of Brain Signals. In: Mana, N., Schwenker, F., Trentin, E. (eds.) ANNPR 2012. LNCS (LNAI), vol. 7477, pp. 225–243. Springer, Heidelberg (2012)
9. Kasabov, N., Dhoble, K., Nuntalid, N., Indiverim, G.: Dynamic Evolving Spiking Neural Networks for On-line Spatio- and Spectro-Temporal Pattern Recognition. Neural Networks 41, 188–201 (2012)
10. Song, S., Miller, K., Abbott, L., et al.: Competitive hebbian learning through spike-timing-dependent synaptic plasticity. Nature Neuroscience 3, 919–926 (2000)
11. Lynch, C.L., Popovic, M.R.: Functional electrical stimulation. Control Systems Magazine, 40–50 (2008)
12. Gerstner, W.: Time structure of the activity of neural network models. Phys. Rev. 51, 738–758 (1995)
13. Delbruck, T.: jAER open source project (2007), `http://jaer.wiki.sourceforge.net`
14. Berry, M.J., Warland, D.K., Meister, M.: The structure and precision of retinal spiketrains. PNAS 94, 5411–5416 (1997)
15. Nuntalid, N., Dhoble, K., Kasabov, N.: EEG Classification with BSA Spike Encoding Algorithm and Evolving Probabilistic Spiking Neural Network. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part I. LNCS, vol. 7062, pp. 451–460. Springer, Heidelberg (2011)

NeuCube Neuromorphic Framework for Spatio-temporal Brain Data and Its Python Implementation

Nathan Scott[1,⋆], Nikola Kasabov[1], and Giacomo Indiveri[2]

[1] Knowledge Engineering and Discovery Research Institute,
Auckland University of Technology, New Zealand
nascott@aut.ac.nz

[2] Neuromophic Cognitive Systems, Institute of Neuroinformatics,
University of Zurich and ETH Zurich, Switzerland

Abstract. Classification and knowledge extraction from complex spatio-temporal brain data such as EEG or fMRI is a complex challenge. A novel architecture named the NeuCube has been established in prior literature to address this. A number of key points in the implementation of this framework, including modular design, extensibility, scalability, the source of the biologically inspired spatial structure, encoding, classification, and visualisation tools must be considered. A Python version of this framework that conforms to these guidelines has been implemented.

Keywords: NeuCube, Neurogenetic, Neuromorphic, Neuroinformatic, Spiking Neural Network, Pattern Recognition.

1 Introduction

The classification and interpretation of spatio-temporal brain data is one of the most complex challenges in modern machine learning. The sheer scale and complexity of the processing units and connections within the human brain lead to highly non-linearly-separable patterns, which make classification of these patterns difficult at best. To this end, a novel computational environment, the NeuCube, has been proposed in prior literature [1]. This paper introduces a general implementation framework, and the specific version written in the Python programming language.

1.1 Overview of the NeuCube Framework

A detailed discussion of the NeuCube Framework or the Spiking Neural Networks upon which it is based is beyond the scope of this paper. For further details see the paper in which it was first proposed [1].

It consists of the following modules: input information encoding module; NeuCube Reservoir module; output module; gene regulatory network (GRN) module.

[⋆] Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 78–84, 2013.

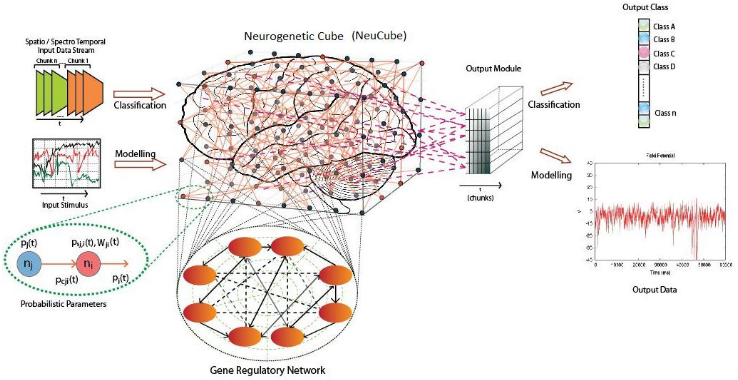


Fig. 1. A block diagram of the NeuCube framework [1]

The input module encodes input data (EEG, fMRI and other brain data) into trains of spikes which are then directly entered into the main module – the NeuCube reservoir (NCR). The NCR consists of Leaky-Integrate-and-Fire Spiking Neurons, a brain inspired computing technique. The framework as a whole (incl. encoding and classification tools) is the 'NeuCube Environment'.

Example Application on Neuroinformatic Data

Training of a NeuCube Architecture: The NC Reservoir is structured to match the spatial distribution of the EEG or fMRI data (or both). This retains the complex spatio-temporal relationships within the data. The available data is converted into spike trains by the encoding module and entered into the corresponding neuron (neurons) in the NC Reservoir. STDP learning is applied to establish the connection weights of spatial-temporal patterns of pathway connectivity. The output classification module (control module) is trained to recognize the states of the NC Reservoir into predefined classes (activate desired control devices).

Recall of the Trained NeuCube Architecture on New Data: New input data is propagated through the NC Reservoir in the same manner as it is initially trained. Output classification (control) results are recorded and the activity of the NeuCube in terms of polychronisation trajectories is analysed and conclusions are made regarding the new input data and the spatio-temporal connectivity and pathways.

Further Adaptation of the NeuCube Architecture: If new data is available that belongs to either existing or new classes, further training of the NeuCube architecture is performed and new output classification (control) neurons are added/evolved and trained in the same way.

2 General Framework for Implementation

The following section describes the general framework for an implementation of the NeuCube system. The italicised headings are aspects that should be considered over and above the standard implementation of the previously discussed theory.

Spatial structure: The spatial positioning for the neurons within this reservoir is drawn directly from the Talairach atlas provided by [2]. Each position in 3D space represents a physical volume within the brain, and has associated Talairach data, including Broadmann areas and the gyri it is conceptually located in. This spatial resolution (number of neurons) should be scaled through the software environment, and should also be capable of mapping an aribitrarily large number of neurons within the volume (reasonably constrained by the computational power available to the environment).

As these locations within the model retain their biological metadata including functional gyri and Broadmann areas, it is possible for us to incorporate location specific connection structures or genetic data based on observed biological phemonmena. For example, neurons in the area representing the occipital lobe could easily be connected in columns and layers, mimicking more accurately the connections in the human brain's occipital lobe.

Connections: These connections are generated homogenously over the whole model, in a small-world manner. The degree and probability of the small-world connections should be configurable. These connections are also capable of performing Spike Time Dependent Plasticity learning [3]. This learning can be adjusted or removed depending on the specific experiment.

Input Mapping: Due to this realistic spatial structure, brain data associated with a physical brain location, can be mapped to the nearest location represented in our NC Reservoir, preserving the complex spatio-temporal relationships within the data.

For example, with EEG data, we use the excellent mapping presented in [4] to associate a standard electrode name location with a specific voxel of brain tissue. See Table 1 for an example of this electrode – location mapping as used in the software environment.

Experiment Configuration: A simple configuration script is generated for each experiment series. This script is written in plain text and contains a number of necessary configuration parameters defining neuron behaviours, connection parameters, simulation times, classification tools, and so on. These files are easily generated in an automated fashion with an included tool, which is of particular use when performing parameter space searches.

Experiments can be run singly or in a batch job form. Multiple experiments can be run on the same data set, including the comparison of different classification tools or encoding techniques.

Table 1. Example locations of EEG data input positions in Talairach space [4], used to select input neurons in the NeuCube Reservoir, retaining the spatio-temporal relationships within the data

Labels	Talairach Coordinates			Gyri	BA
	x	y	z		
F7	-52.1 ± 3.0	28.6 ± 6.4	3.8 ± 5.6	L FL Inferior Frontal	45
T7	-65.8 ± 3.3	17.8 ± 6.8	-2.9 ± 6.1	L TL Middle Temporal	21

Spike Encoding: A number of different spike encoding schemes should be provided, as different forms of input data require different types of preprocessing. These take input and provide an output in a standard form across all encoding schemes. Examples can be found in [1].

Classification: The implementation of multiple classification tools is necessary, as these serve different end goals for the system. Particular examples include SPAN [5] for motor signal control and deSNN variants [6] for classification tasks.

Statistics: A statistics and monitoring package, responsible primarily for the calculation of statistics related to classification error rates and connection parameters, is necessary.

Parallelisation and Clustering: Scalability is handled through the use of local machine parallelisation and multi-machine clustering. A standard protocol such as MPI is encouraged.

Extensibility: The environment should be written in such a way that it is easily extended with a common language. Functionality that may be extended includes neural models (for example, implementing a probabilistic neural model as [7]), connection models, or the addition of tools such as a neurogenetic optimisation module [1].

2.1 Use of the Tools Independently

As this software environment is developed in a modular fashion, each of these modules can be used independently or in concert depending on the user's requirements. For example, a data sample can be encoded into spikes using the AER Encoding module, and then presented directly to the deSNN module without using the reservoir. Similarly, a pre-encoded spike train can be presented to the NeuCube Reservoir and the outputs recorded and used elsewhere. See Table 2 for the independent modules currently included in the environment. The main categories of these are briefly discussed below.

Encoding Module: The included spike train encoding tools can be used to convert a number of types of spatio-temporal data into spike trains suitable for general use in either the PyNEST or Brian simulators.

Table 2. Software tools within the Python NeuCube Environment capable of being used in a modular fashion either within or external to the environment

Category	Type
Encoding	AER BSA Population
Reservoir	NeuCube (STDP) NeuCube (static synapses) Standard LSM
Classification	MultiSPAN sSPAN deSNNs deSNNr
Utilities	SNN Visualisation Experiment file generator Statistics

NeuCube Reservoir Module: The NC Reservoir module can be used independently as a standard SNN reservoir. An option is provided to disable STDP learning. In addition, this implementation has the capacity to include a standard LSM in place of the NCR, to allow for non-NI data to be modelled.

Classification Module: The classification techniques implemented can be applied to a wide range of spiking neural network pattern recognition tasks, independent of the NC module.

3 Python Version

An implementation of this framework has been developed in the Python programming language, using the NEST spiking neural network simulator's PyNEST interface. This implementation has been created following the guidelines previously established in this paper, with particular emphasis on modularity and extensibility.

Extensibility: This implementation (including the PyNEST simulator) is easily extended through standard Python and C++. Computationally intensive tasks are vectorised or moved to C++.

Visualisation Tools: Of specific interest is the capacity to visualise both spiking activity and the evolution of neural connectivity over the life cycle of the simulation, in 3D. This is particularly useful for knowledge extraction from the model, and in itself represents a novel contribution to the field.

Standard SNN visualisation tools including raster plots of spiking activity are included, and can be extended easily.

Hardware Implementations: The Python software implementation will allow direct neuromorphic implementations on SNN chips and systems as follows:

1. Implementation of the NeuCube framework on a SNN SRAM chip [8,9].
2. Implementation on the SpiNNaker SNN supercomputer [10]. A Python version of the NC is desirable as the SpiNNaker project is compatible with PyNN scripts [11], providing the NC environment with the capacity to be run on this large-scale, dedicated hardware system.
3. Implementation on a memristor based highly-parallel computation system currently in development [12] is also theoretically possible, and warrants exploration.

3.1 Future Additions

A number of extensions are planned for future versions of this environment. The primary two are mentioned briefly below.

Neurogenetic Optimisation Module: The neurogenetic optimisation module described in [1] will be implemented. This module takes real genetic data acquired from the Allen Brain Atlas [13] associated with the spatial location of the neurons in the NC Reservoir and uses this to modulate the behaviour of these neurons and their connectome. It is also possible to use this gene and brain composition data to structure the simulation's connections in such a way that they are more biologically plausible.

Multi-Simulator Support: As aspects of the model are already compatible with both the PyNEST and Brian SNN simulators, it is a trivial matter to add the necessary functionality to make this environment fully compatible with Brian.

The feasibility of implementing this environment in PyNN will be explored in the near future. This would provide implicit multi-simulator support [14].

4 Conclusion

This paper has presented an introduction to a new spiking neural network brain data classification framework, and a specific Python implementation of the same.

The environment is also capable of application on general spatio- or spectro-temporal data. As it is implemented in a modular fashion, selected components can be utilised separately or as a whole depending on specific user needs. Simple configuration of experiments through a standard text file format allows for automated generation and execution of large numbers of experiments. Large experiments completed at high speeds are possible through the use of local multithreading, and clustering via MPI. The code is easily extensible, and visualisation tools allow for novel knowledge extraction from the system's dynamic behaviour.

References

1. Kasabov, N.: NeuCube EvoSpike Architecture for Spatio-Temporal Modelling and Pattern Recognition of Brain Signals. In: Mana, N., Schwenker, F., Trentin, E. (eds.) ANNPR 2012. LNCS (LNAI), vol. 7477, pp. 225–243. Springer, Heidelberg (2012)
2. Lancaster, J.L., Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., Kochunov, P.V., Nickerson, D., Mikiten, S.A., Fox, P.T.: Automated Talairach Atlas labels for functional brain mapping. Human Brain Mapping 10, 120–131 (2000)
3. Kepecs, A., van Rossum, M.C.W., Song, S., Tegner, J.: Spike-timing-dependent plasticity: common themes and divergent vistas. Biological Cybernetics 87, 446–458 (2002)
4. Koessler, L., Maillard, L., Benhadid, A., Vignal, J.P., Felblinger, J., Vespignani, H., Braun, M.: Automated cortical projection of EEG sensors: Anatomical correlation via the international 10-10 system. NeuroImage 46, 64–72 (2009)
5. Mohemmed, A., Schliebs, S., Matsuda, S., Kasabov, N.: SPAN: Spike Pattern Association Neuron for Learning Spatio-Temporal Sequences. Int. J. of Neural Systems 22(4), 1–16 (2012)
6. Kasabov, N., Dhoble, K., Nuntalid, N., Indiveri, G.: Dynamic Evolving Spiking Neural Networks for On-line Spatio- and Spectro-Temporal Pattern Recognition. Neural Networks 41, 188–201 (2013)
7. Kasabov, N.: To spike or not to spike: A probabilistic spiking neuron model. Neur. Netw. 23(1), 16–19 (2010)
8. Indiveri, G., Stefanini, F., Chicca, E.: Spike-based learning with a generalized integrate and fire silicon neuron. In: 2010 IEEE Int. Symp. Circuits and Syst., Paris, pp. 1951–1954 (2010)
9. Indiveri, G., Horiuchi, T.K.: Frontiers in neuromorphic engineering. Frontiers in Neuroscience 5, 1–2 (2011)
10. Furber, S.: To Build a Brain. IEEE Spectrum 49(8), 39–41 (2012)
11. Galluppi, F., Rast, A., Davies, S., Furber, S.: A general-purpose model translation system for a universal neural chip. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (eds.) ICONIP 2010, Part I. LNCS, vol. 6443, pp. 58–65. Springer, Heidelberg (2010)
12. Serrano-Gotarredona, T., Prodromakis, T., Indiveri, G., Linares-Barranco, B., Masquelier, T.: STDP and STDP variations with memristors for spiking neuromorphic learning systems. Frontiers in Neuroscience 7 (2013)
13. Hawrylycz, M.J., et al.: An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489, 391–399 (2012)
14. Davison, A.P., Brüderle, D., Eppler, J.M., Kremkow, J., Muller, E., Pecevski, D.A., Perrinet, L., Yger, P.: PyNN: A common interface for neuronal network simulators. Frontiers in Neuroinformatics 2(11), 1–10 (2008)

Vector Quantization Using Mixture of Epsilon-Insensitive Components

Kazuho Watanabe

Graduate School of Information Science, Nara Institute of Science and Technology,
8916–5, Takayama-cho, Ikoma, Nara, 630–0192 Japan
wkazuho@is.naist.jp

Abstract. We consider mixture models consisting of ε-insensitive component distributions, which provide an extension of Laplacian mixture models. An EM-type learning algorithm is derived for maximum likelihood estimation of the mixture models. The derived algorithm is applied to approximate computation of rate-distortion functions associated with the ε-insensitive loss function. Then the robustness property of the mixture of ε-insensitive component distributions is demonstrated in a multi-dimensional mixture modelling problem.

1 Introduction

Mixture models are widely used for clustering, quantization and density estimation. In particular, Laplacian mixture models have been proposed and applied for the purposes of robust clustering and overcomplete source separation [4,8]. In this article, we consider an extension of the Laplacian mixture model to the mixture of ε-insensitive component distributions. The ε-insensitive distribution is defined by an ε-insensitive loss function which, when $\varepsilon = 0$, corresponds to the absolute loss function appearing in the Laplace distribution. The ε-insensitive loss function has been used in the support vector regression and other related methods to provide a sparsity inducing mechanism [3,5,9,10,11]. In a previous work, upper and lower bounds were obtained for the rate-distortion function associated with the ε-insensitive loss function [12]. Although the rate-distortion function shows the theoretically optimal performance of quantization schemes using the ε-insensitive loss function as a distortion measure, its explicit evaluation has yet to be obtained, and the optimal reconstruction distribution achieving the rate-distortion function is still unknown.

In this article, we derive an Expectation-Maximization (EM)-type learning algorithm for maximum likelihood estimation of mixtures of ε-insensitive component distributions, which provides an extension of the algorithm for Laplacian mixture models [8]. We apply it to 1-dimensional problems where the rate-distortion functions associated with the ε-insensitive distortion measure are approximately computed. We also examine the convergence property of the learning algorithm numerically. Then we apply the derived algorithm to a multi-dimensional data set in order to demonstrate the robustness-enhancing feature of the ε-insensitive component distribution.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 85–92, 2013.

2 Mixture of ε-Insensitive Component Distribution

For $x \in \mathbf{R}^d$, let

$$p(x|w) = \sum_{k=1}^{K} a_k c_\varepsilon(x|\theta_k) \tag{1}$$

be the mixture model of the component distribution $c_\varepsilon(x|\theta)$. The parameter vector w consists of the parameter $\theta_k \in \mathbf{R}^d$ for each component and the mixing proportions $\{a_k\}$ satisfying $a_k \geq 0$ for $k = 1, 2, \cdots, K$ and $\sum_{k=1}^{K} a_k = 1$.

In this paper, we focus on the following component distribution:

$$c_\varepsilon(x|\theta) = \frac{1}{C_s} \exp\left\{-s\rho_\varepsilon(||x-\theta||)\right\}, \tag{2}$$

defined by the ε-insensitive loss function $\rho_\varepsilon(z) = \max\{|z| - \varepsilon, 0\}$ and the Euclidean distance between x and θ, $||x - \theta|| = \sqrt{\sum_{j=1}^{d}(x_j - \theta_j)^2}$. s is a positive real (global) parameter, which can also be included in the component parameter. In Eq. (2), the normalization constant C_s is explicitly obtained as

$$\begin{aligned} C_s &= \int_{x \in \mathbf{R}^d} \exp\left\{-s\rho_\varepsilon(||x||)\right\} dx = I(d) \int_0^\infty e^{-s\rho_\varepsilon(r)} r^{d-1} dr \\ &= I(d) \left\{ \frac{\varepsilon^d}{d} + \frac{e^{s\varepsilon}}{s^d} \Gamma(d, s\varepsilon) \right\}, \end{aligned}$$

where $I(d) = \frac{d\sqrt{\pi}^d}{\Gamma(d/2+1)}$ is the area of the d-dimensional unit hypersphere and $\Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt$ and $\Gamma(u, \alpha) = \int_\alpha^\infty t^{u-1} e^{-t} dt$ are the gamma and the upper incomplete gamma functions respectively.

When $\varepsilon = 0$, the component (2) reduces to the (isotropic) Laplace distribution, $c_0(x|\theta) \propto \exp(-s||x - \theta||)$, and the mixture (1) reduces to the Laplacian mixture model [7,4,8]. We refer to the mixture model in Eq. (1) as the ε-insensitive mixture model (EIMM).

3 EM Algorithm for EIMM

We derive a learning algorithm for maximizing the likelihood of the EIMM based on the EM algorithm [6,1].

3.1 E and M Steps

Given training samples $x^n = \{x_1, \cdots, x_n\}$, the log-likelihood of the EIMM is lower bounded as follows,

$$\begin{aligned} &\sum_{i=1}^{n} \log p(x_i|w) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} a_k c_\varepsilon(x_i|\theta_k) \\ &\geq \sum_{i=1}^{n} \sum_{k=1}^{K} \eta_{ik} \left\{ \log a_k - \log C_s - s\rho_\varepsilon(||x_i - \theta_k||) - \log \eta_{ik} \right\} \equiv Q(w|\tilde{w}) \end{aligned}$$

where

$$\eta_{ik} = \frac{\tilde{a}_k c_\varepsilon(x_i|\tilde{\theta}_k)}{\sum_{l=1}^{K} \tilde{a}_l c_\varepsilon(x_i|\tilde{\theta}_l)} \tag{3}$$

is the responsibility representing the probability that x_i is assigned to the kth component under the current estimate of the model parameter $\tilde{w} = \{\tilde{a}_k, \tilde{\theta}_k\}$, and is satisfying $\sum_{k=1}^{K} \eta_{ik} = 1$ for $i = 1, \cdots, n$. Maximizing $Q(w|\tilde{w})$ with respect to w leads to the EM algorithm, which sets an initial value for $\tilde{w}$ and iterates the following E- and M-steps until convergence, and is guaranteed to increase the log-likelihood at each iteration:
E-step: Compute η_{ik} for $i = 1, \cdots, n$ and $k = 1, \cdots, K$ by Eq. (3).
M-step: For $k = 1, \cdots, K$,

$$\tilde{a}_k \leftarrow \frac{1}{n}\sum_{i=1}^{n} \eta_{ik}, \quad \tilde{\theta}_k \leftarrow \underset{\theta_k}{\operatorname{argmin}} \sum_{i=1}^{n} \eta_{ik}\rho_\varepsilon(||x_i - \theta_k||). \tag{4}$$

Note that unlike for the Gaussian mixture model, the updating rule of θ_k in the M-step is not explicitly solved. We focus on the minimization problem (4) in the next subsection.

In order to estimate the parameter s, we can use the first order approximation, $\log C_s \simeq \log \frac{I(d)\Gamma(d)}{s^d} + s\varepsilon$, and include the update rule,

$$\frac{1}{\tilde{s}} \leftarrow \frac{1}{d}\sum_{i=1}^{n}\sum_{k=1}^{K} \eta_{ik}\rho_\varepsilon(||x_i - \theta_k||) + \frac{\varepsilon}{d}.$$

3.2 Dual Problem for M-Step and Partial M-Step

The M-step of the EM algorithm requires minimizing a convex function of the form,

$$L(\theta) = \sum_{i=1}^{n} \nu_i \rho_\varepsilon(||x_i - \theta||), \tag{5}$$

where $0 \leq \nu_i \leq 1$ for $i = 1, \cdots, n$. By introducing slack variables $\xi = \{\xi_i\}_{i=1}^{n}$, minimization of (5) is reformulated as the following minimization problem with inequality constraints:

$$\min_{\theta,\xi} \sum_{i=1}^{n} \nu_i \xi_i, \text{ subj. to } ||x_i - \theta|| - \varepsilon \geq \xi_i \text{ and } \xi_i \geq 0 \;\; (i = 1, \cdots, n).$$

Through the Lagrange dual problem of the above minimization problem we can see that

$$L(\theta) = \max_{\alpha \in B} \tilde{L}(\alpha, \theta), \tag{6}$$

where $\alpha = (\alpha_1, \cdots, \alpha_n)$, $\tilde{L}(\alpha, \theta) = \sum_{i=1}^{n} \alpha_i(||x_i - \theta|| - \varepsilon)$, and $B = \{(\alpha_1, \cdots, \alpha_n) : 0 \leq \alpha_i \leq \nu_i, i = 1, \cdots, n\}$. In fact, the maximum with respect to α_i is achieved when

$$\alpha_i = \begin{cases} \nu_i \;\; (||x_i - \theta|| > \varepsilon) \\ 0 \;\; (||x_i - \theta|| \leq \varepsilon) \end{cases} \tag{7}$$

for $i = 1, \cdots, n$. Putting this back into (6) yields the original form of $L(\theta)$ in Eq. (5).

If we first minimize with respect to θ instead of maximizing with respect to α in (6), we set the derivative of $\tilde{L}(\alpha, \theta)$ to zero,

$$\frac{\partial \tilde{L}}{\partial \theta} = \sum_{i=1}^{n} \alpha_i \frac{(\theta - x_i)}{||x_i - \theta||} = 0 \quad \Rightarrow \quad \theta = \frac{\sum_{i=1}^{n} \frac{\alpha_i}{||x_i - \theta||} x_i}{\sum_{i=1}^{n} \frac{\alpha_i}{||x_i - \theta||}} \tag{8}$$

Hence, we can think of the fixed-point optimization approach that iterates (7) and (8) to solve the minimization of $L(\theta)$. However, this approach can fail to minimize $L(\theta)$ although it does fully minimize $L(\theta)$ in some cases as we will partly see in Section 4. Instead we propose a single iteration procedure which iterates (7) and (8) once at each M-step. While this procedure does not fully minimize $L(\theta)$, if the updating rule (8) decreases $L(\theta)$ even a little, the overall EM algorithm monotonically increases the likelihood. This is an example of the so-called "partial M-step" [1], and reduces to the learning algorithm of Laplacian mixture model proposed in [8] when $\varepsilon = 0$. Algorithmically, introducing $\varepsilon > 0$ stabilizes the algorithm of [8] which can be unstable when $||x_i - \theta||$ takes a value close to zero (see Eq. (8)). This is because, when $\varepsilon > 0$, Eq. (7) sets $\alpha_i = 0$ if $||x_i - \theta||$ is small enough.

4 Application to Rate-Distortion Computation

In this section, we apply the learning algorithm developed in the previous sections to approximate computation of the rate-distortion function for the ε-insensitive loss [2,12]. The rate-distortion function is obtained by minimizing the mutual information subject to an average distortion constraint. This problem can be reformulated as a problem of minimizing the following functional over the output (reconstruction) density $q(\theta)$ [2,12]:

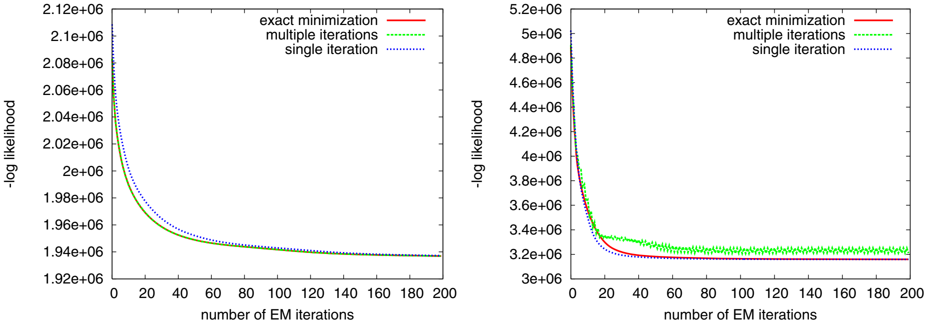
$$F(q) = -\int p(x) \left[\log \int e^{-sd(x,\theta)} q(\theta) d\theta \right] dx, \tag{9}$$

where $p(x)$ is the density of the source, and $d(x, \theta)$ is the distortion measure between x and θ. If $\hat{q}(\theta)$ is the optimal output density, then the optimal conditional output density is given by $\hat{q}(\theta|x) \propto \hat{q}(\theta) \exp(-sd(x, \theta))$. The rate and average distortion corresponding to the slope parameter s are

$$R(D_s) = \int p(x) \hat{q}(\theta|x) \log \frac{\hat{q}(\theta|x)}{\int p(\tilde{x}) \hat{q}(\theta|\tilde{x}) d\tilde{x}} d\theta dx \tag{10}$$

$$D_s = \int p(x) \hat{q}(\theta|x) d(x, \theta) dx d\theta \tag{11}$$

$-s$ provides the slope of the tangent of the rate-distortion function $R(D)$ at $(D_s, R(D_s))$.



(a) $K = 10$, $s = 5$ (b) $K = 10$, $s = 25$

Fig. 1. Evolution of the negative log-likelihood against EM iterations for (a) $K = 10$, $s = 5$ and (b) $K = 10$, $s = 25$

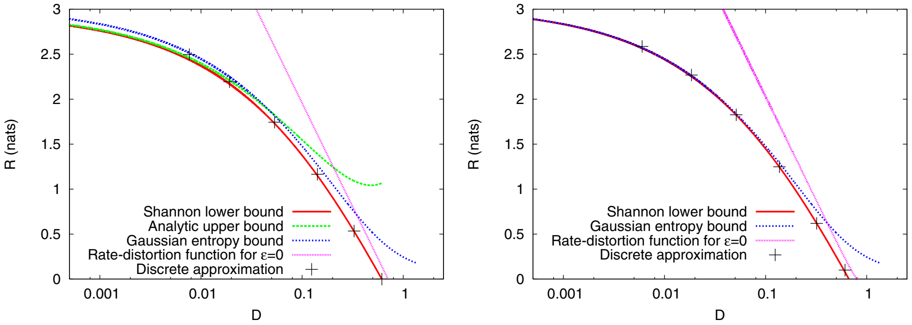
If we take $d(x, \theta) = \rho_\varepsilon(||x - \theta||)$, the above problem of minimizing (9) reduces to minimizing the KL-divergence form $p(x)$ to the mixture of ε-insensitive distributions (2) mixed by $q(\theta)$. Hence, this problem is approximated by the maximum likelihood of the model $\int q(\theta) c_\varepsilon(x|\theta) d\theta$ if we approximate the source $p(x)$ by the empirical distribution, $\hat{p}(x) = \sum_{i=1}^{n} \delta(x - x_i)$, where δ is Dirac's delta function, of the samples $\{x_1, \cdots, x_n\}$ drawn i.i.d. from $p(x)$. Here, we further restrict the reconstruction density $q(\theta)$ to be a K-component discrete distribution, $q(\theta) = \sum_{k=1}^{K} a_k \delta(\theta - \theta_k)$. Then the rate-distortion function is finally approximated by obtaining the maximum likelihood estimate $\hat{w}$ for the parameter of the mixture of ε-insensitive distributions (1) for each slope parameter s.

We focused on the 1-dimensional case, $d = 1$, and fixed $\varepsilon = 0.1$ throughout the experiment. We generated two data sets of size $n = 10^6$ according to the standard normal distribution and the Laplace distribution with the density $l_\beta(x) = \frac{\beta}{2} e^{-\beta|x|}$ $(\beta = 1/\sqrt{2})$ respectively.

We first examined the convergence property of the iterative procedure developed in Section 3.2. The golden section search was applied for solving the minimization of $L(\theta)$ in Eq. (5) exactly. The M-step using this exact minimization procedure is refered to as "exact minimization." We refer to the M-step that iterates Eqs. (7) and (8) multiple times (up to 200 times) as "multiple iterations" and the M-step that iterates them once as "single iteration."

In most cases, the multiple iteration minimized the loss function (5) in every M-step and the overall evolution of the negative log-likelihood against EM iterations coincided with that of the exact minimization as demonstrated for the case of $K = 10$, $s = 5$ and the Gaussian data set (Fig.1(a)). However, the multiple iterations can fail to minimize the loss function (5) for example when there is a severe mismatch in K (or s) as demonstrated for the case of $K = 10$ and $s = 25$ and the Gaussian data set (Fig.1(b)). As implied from these figures, we can detect this by monitoring the monotonicity of the likelihood. On the other hand, the EM algorithm using the single iteration monotonically decreased the negative log-likelihood in both cases (Fig.1(a) and Fig.1(b)). It does not fully minimizes the loss function in each M-step, the convergence speed of its overall EM algorithm can be slower than that of the EM algorithm using exact minimization while these two both converged to the same estimate (Fig.1(a)). The EM algorithm using the single iteration M-step can be faster than the EM algorithm using exact minimization (Fig.1(b)).

Next, applying the EM algorithm with the exact minimization M-step, we approximately calculated the 6 points on the rate-distortion curve corresponding to $s = 1.25, 2.5, 5, 10, 20, 40$. For each s, we applied the EIMM with $K = 2, 4, \cdots, 48, 50$ and adopted the number of components K when the increase in the likelihood was saturated. We calculated the resulting rate (10) and average distortion (11) for the two data sets, the Laplacian data set (Fig.2(a)) and the Gaussian data set (Fig.2(b)). Also plotted in these figures are the upper and lower bounds for the rate-distortion curve which was obtained in [12]. For the both data sets, the pairs of rate and distortion for $s = 1.25, 2.5, 5, 10$ are located between the upper and lower bounds and are very close to the Shannon lower bound, which was proved to be strictly smaller than the exact rate-distortion curve for all D [12]. This implies that the Shannon lower bound provides a very accurate approximation to the exact rate-distortion curve and that the optimal reconstruction distribution can be well approximated by a discrete distribution. The points for $s = 40$ (for the Laplacian data set) and $s = 20, 40$ (for the Gaussian data set) are located above the upper bounds. This seems due to the limited number of mixture components (up to 50) and the limited number of EM iterations (up to 500 iterations).



(a) Laplacian source

(b) Gaussian source

Fig. 2. Rate-distortion bounds (curves) and approximated values of rate-distortion pairs (crosses) for (a) the Laplacian data set and (b) the Gaussian data set. Only the lowest curve in each panel is a lower bound, while the remaining curves (or lines) are upper bounds.

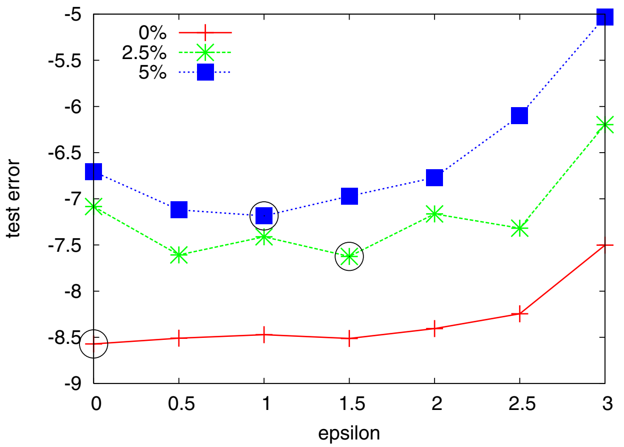


Fig. 3. Average test errors for different ε. The minimum for each contamination level is marked by a circle. The minimums of the average test errors for the contamination levels, 2.5% and 5% are significantly smaller than those of $\varepsilon = 0$ (paired t-test, $p < 0.05$).

5 Application to Multi-dimensional Problem

It was demonstrated for the support vector regression that the ε-insensitive loss function induces robustness [3,5,9,10,11]. We investigate the robustness property of EIMMs by using 10-dimensional synthetic data set.

We generated 500 samples $\{x_i\}_{i=1}^{500}$ from a 5-component isotropic Laplacian mixture model (LMM) on 10-dimensional space. The mean parameters of the true LMM were fixed to points randomly generated from the uniform distribution on $[-5,5]^{10}$ and we set $s = 5$. As a contamination, we replaced $C = 0$, 2.5 and 5% of data by random points uniformly distributed on $[-5,5]^{10}$ and made 3 data sets. We applied the EM algorithm using the partial M-step for the EIMMs with $\varepsilon = 0$ (LMM), 0.5, 1, 1.5, 2, 2.5 and 3 and obtained the estimate $\hat{w} = \{\hat{a_k}, \hat{\theta}_k\}$ for each EIMM. We generated the test data $\{\tilde{x}_i\}_{i=1}^{T}$ ($T = 25000$) from the true LMM (without contamination) and calculated the test error measured by the negative log-likelihood,

$$E(x^n) = -\frac{1}{T}\sum_{i=1}^{T}\log\sum_{k=1}^{K}\hat{a}_k c_0(\tilde{x}_i|\hat{\theta}_k),$$

where we set $\varepsilon = 0$ to ignore the influence of model mismatch and compare the accuracy of estimates for different ε. We repeated the experiment 100 times using different training data sets obtained from the same generation process and calculated the average of the test errors (Fig.3). It can be seen that introducing a positive ε reduces the average test error when there is a contamination. This implies that robustness is enhanced by the ε-insensitive component distribution.

6 Conclusion

In this study, we derived an EM-type algorithm for the EIMM. As demonstrated in Section 4 for 1-dimensional problems, a 1-dimensional search technique such as the golden section search is applicable to the M-step. For higher-dimensional problems, however, this is not the case. Alternatively, we can use the partial M-step proposed in Section 3.2 which executes a single iteration of Eqs. (7) and (8). It is an important undertaking to investigate the convergence property of this EM-type algorithm in higher-dimensional problems. Higher-dimensional extensions of the rate-distortion analysis are also to be addressed.

References

1. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press (2012)
2. Berger, T.: Rate Distortion Theory: A Mathematical Basis for Data Compression. Prentice-Hall, Englewood Cliffs (1971)
3. Chu, W., Keerthi, S.S., Ong, C.J.: A unified loss function in Bayesian framework for support vector regression. In: Proc. of ICML, pp. 51–58 (2001)
4. Cord, A., Ambroise, C., Cocquerez, J.: Feature selection in robust clustering based on Laplace mixture. Pattern Recognition Letters 27(6), 627–635 (2006)
5. Dekel, O., Shalev-Shwartz, S., Singer, Y.: ε-insensitive regression by loss symmetrization. Journal of Machine Learning Research 6, 711–741 (2005)
6. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39-B, 1–38 (1977)
7. Eltoft, T., Kim, T., Lee, T.: On the multivariate Laplace distribution. IEEE Signal Processing Letters 13(5), 300–303 (2006)
8. Mitianoudis, N., Stathaki, T.: Batch and online underdetermined source separation using Laplacian mixture models. IEEE Transactions on Audio, Speech and Language Processing 15(6), 1818–1832 (2007)
9. Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L.: New support vector algorithms. Neural Computation 12(5), 1207–1245 (2000)
10. Steinwart, I., Christmann, A.: Support Vector Machines. Springer (2008)
11. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1995)
12. Watanabe, K.: Rate-distortion bounds for an ϵ-insensitive distortion measure. In: Proc. of ITW 2013, pp. 679–683 (2013)

Applicability of ICA-Based Dimension Reduction in Fuzzy *c*-Means-Based Classifier

Takuya Kobayashi[1], Katsuhiro Honda[1], Akira Notsu[1],
and Hidetomo Ichihashi[2]

[1] Department of Computer Science and Intelligent Systems
Osaka Prefecture University
1-1 Gakuen-cho, Nakaku, Sakai, Osaka 599-8531, Japan
{honda,notsu}@cs.osakafu-u.ac.jp
[2] Faculty of Economics
Osaka University of Economics and Law
6-10 Gakuonji, Yao, Osaka 581-8511, Japan
ichi@cs.osakafu-u.ac.jp

Abstract. Fuzzy *c*-Means-based Classifier (FCMC) has been proved to have high performances based on clustering concepts in conjunction with several parameter optimization methods. In general, FCMC is applied to high-dimensional data after dimension reduction by Principal Component Analysis (PCA). In this paper, the applicability of Independent Component Analysis (ICA)-based dimension reduction is investigated in the FCMC context. ICA is a computational method for separating a multivariate signal into additive subcomponents with the assumption of non-Gaussian signals. This paper compares the performance of FCMC using four data sets. Two initialization approaches of the PCA-Tree-based and k-dimensional tree (kd-Tree)-based are also compared.

Keywords: Classifier, Clustering, Principal component analysis, Independent component analysis.

1 Introduction

Fuzzy *c*-Means-based Classifier (FCMC) is a simple pattern classification approach based on the clustering concept and the model parameters are optimized by several heuristic approaches. FCMC shows a high classification performance on high-dimensional data sets, and has been proved to have advantages compared to LibSVM [1]. In the past study, the original feature dimensions of the data are reduced by Principal Component Analysis (PCA), but other compression method has not been tested.

Independent Component Analysis (ICA) is an unsupervised technique, which in many cases characterizes data in a natural way, and is a useful technique for Projection Pursuit as well [2] . In the general formulation of ICA, the purpose is to transform an observed vector linearly into the vector whose components are statistically as independent from each other as possible. The mutual dependence

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 93–100, 2013.

of the components is classically measured by their non-Gaussianity. Maximizing the non-Gaussianity gives us one of the independent components. Therefore, the basis vectors of ICA should be especially useful in Projection Pursuit and in extracting characteristic features from natural data.

In this paper, the applicability of the ICA-based dimension reduction instead of PCA-based one is investigated through several comparative experiments using four benchmark data sets, which are available from the UCI benchmark repository. In the experiments, two ways of initializing cluster centers are also compared. In FCMC, the following two approaches are applicable, 1) PCA-Tree [3]: the splitting hyper-plane is perpendicular to the first PCA basis vector of each internal node cluster, and 2) kd-tree [4,5]: the splitting hyper-plane is perpendicular to an original coordinate axis.

2 Fuzzy c-Means-Based Classifier

In FCMC, the membership function of a modified type of FCM clustering [6] is used for classification. The first phase of FCMC is the generalized hard clustering [7,8], in which FCM-type clustering is performed in a defuzzified manner. The objective function of FCM with regularization by Kullback-Leibler divergence (KLFCM) [8,9] is linearized and the update rules are derived by using Lagrangian multiplier method. Let $x_k \in \mathcal{R}^p$ be a feature vector and u_{ki} be the cluster indicator of membership value of x_k in the i-th cluster, which is estimated by nearest cluster allocation. The clustering criterion is given by the squared Mahalanobis distance from x_k to cluster center $v_i \in \mathcal{R}^p$:

$$D(x_k, v_i; S_i) = (x_k - v_i)^\top S_i^{-1}(x_k - v_i). \tag{1}$$

S_i is a covariance matrix of data samples of the i-th cluster. Let the mixing proportion of i-th cluster be

$$\alpha_i = \frac{\sum_{k=1}^{N} u_{ki}}{\sum_{j=1}^{c}\sum_{k=1}^{N} u_{kj}} = \frac{1}{N}\sum_{k=1}^{N} u_{ki}, \tag{2}$$

where c denotes the number of clusters and N denotes the number of samples. The updating rule is called as the generalized hard c-means [7,8].

Although initial locations of cluster centers or membership values are usually given randomly in the FCM clustering, the classification performance is severely sensitive to initialization. In order to obtain stable performances, FCMC adopts the bisection method based on PCA-Tree or kd-Tree. For example in PCA-Tree, let $f_k, k = 1, ..., N$ be PCA scores of the data set $X = (x_1, ..., x_N)^\top$ of a class. f_k are associated with the largest singular value of mean corrected X. Initial memberships of the 1st cluster are given to the data with positive f_k. Those of the 2nd cluster are given to the data with negative f_k. This bisection procedure is repeated on each cluster until the number of clusters becomes equal to a prespecified number $c = 2^h$, where h is the height of a complete binary tree.

One of the impediments for FCMC is the singularity of covariance matrices, which frequently occurs when feature dimension is relatively high and the number of samples in each cluster is small. So, the low-rank approximation of covariance matrices in the mixture of probabilistic principal component analysis (MPCA) [10] and the character recognition [11] is applied. By reducing the number r of basis vectors in the approximation of S_i, the algorithm becomes stably convergent. When r=0, S_i is a diagonal matrix, hence is non-singular, and $D(x_k, v_j; S_j)$ is reduced to Euclidean distance.

The clustering is done on a per class bases. The classification (i.e., the second phase) is performed by computing fuzzy memberships. Let π_q denote the mixing proportion (i.e., a priori probability) of class q. Let α_{qj} be α_i in (2) for cluster j of class q. The class membership of k-th data x_k to class q is computed as:

$$u^*_{qjk} = \alpha_{qj}|S_{qj}|^{-\frac{1}{\gamma}}(D(x_k, v_{qj}; S_{qj}) + \nu)^{-\frac{1}{m}}, \tag{3}$$

$$\tilde{u}_{qk} = \frac{\pi_q \sum_{j=1}^{c} u^*_{qjk}}{\sum_{s=1}^{Q} \pi_s \sum_{j=1}^{c} u^*_{sjk}}, \tag{4}$$

where c denotes the number of clusters of each class and Q denotes the number of classes. We selected the functional form of u^* based on the membership functions derived from the generalized FCM objective function [8] and that of FCM with regularization by Kullback-Leibler divergence [8,9].

At the completion of clustering for all classes in the first phase of FCMC, we compute Mahalanobis distances by (1) for all the samples in the test set and then the distances are fixed. This distance calculation part is coded in Visual C in the revised training program [1].

The second phase of FCMC is the parameter optimization and the hyperparameters, i.e., $m \in [0, 2], \gamma \in [0, 20]$ are selected to minimize error rate on the test sets. $\nu = 5$ is fixed for all the benchmark data in this paper. These ranges or intervals are fixed and used for all the data sets in [12,13] and also for all the data sets used in this paper.

3 ICA Formulation and Fast ICA Algorithm

In the conventional research, we applied FCMC to high-dimensional data after PCA-based dimension reduction and have demonstrated high performances [12,13]. In this research, the applicability of ICA-based dimension reduction is investigated.

Let v and s be M-dimensional observed data vector and N $(N \leq M)$ dimensional source signal vector, respectively. In the ICA formulation, v is assumed to be the linear mixture of s_i as follows:

$$v = As, \tag{5}$$

where the elements of source signals $(s_1, s_2, ..., s_D)$ are mutually statistically independent and have zero-means. The unknown $K \times D$ matrix A is called the mixing matrix to be reconstructed and the goal of ICA is to estimate the source signals s_i and the mixing matrix A using only the observed data v.

Fast ICA algorithm proposed by Hyvärinen *et al.* [14] is a useful algorithm that is very simple and fast to converge. Generally, in a PCA-based preprocessing, observed data v are transformed into linear combinations z,

$$z = Mv = MAs = Bs, \tag{6}$$

such that its elements $(z_1, z_2, ..., z_D)$ are mutually uncorrelated and all have unit variance, and $B = MA$ is an orthogonal matrix. The elements of B is derived by minimizing or maximizing the following objective function:

$$J(w_i) = E((w_i^\top x)^4) - 3||w_i||^4 + F(||w_i||^2), \tag{7}$$

where w_i corresponds to one of the columns of the mixing matrix B. The first two terms represent the fourth-order cumulant or kurtosis for measuring non-Gaussianity to be maximizing through the fixed-point algorithm for ICA.

In this paper, we downloaded the source code from http://research.ics.aalto.fi/ica/fastica, and implemented Fast ICA.

4 Experiments

In this section, we report the classification performance of FCMC on four benchmark data sets available from the UCI benchmark repository. Table 1 summarizes the characteristics of the benchmark data sets used in this paper. The original feature dimensions are shown in the column "feature dimensions". The numbers of data samples are given in column "training data" and "testing data".

Table 1. Benchmark data

data name	feature dimensions	training data	testing data
Heart	44	80	187
Iono	33	200	150
Sonar	60	104	104
Wine	12	2000	4497

In the previous study, following two ways of initial partitioning in the clustering step of FCMC are compared [1]. 1) PCA-Tree: the splitting hyper-plane is perpendicular to the first PCA basis vector of each internal node cluster, and 2) kd-Tree: the splitting hyper-plane is perpendicular to an original coordinate axis. So this paper reports the results of comparing between ICA and PCA in FCMC using the above four benchmark datasets, where the two initialization methods of clustering are used. Experiments are performed on Dell Precision T3500, 2.67GHz 3.25GB, Windows XP.

4.1 Herat, Iono, and Sonar Data

Tables 2-4 show the classification performance on three data sets (Herat, Iono, and Sonar data). In Heart data (5 dimension, PCA-tree) and Iono data (5 dimension, PCA-tree), the test error rates of ICA were worse than those of PCA. But in other cases, the test error rates of ICA were better than those of PCA. And the test error rates of FCMC is nearly the same with those of other classifiers [15,16,17]. For example, the CLIP3 machine learning algorithm achieved 77.0% accuracy on Heart data [15]. So ICA-based approach is effective for FCM classifier on those data sets.

Table 2. Comparison of accuracy on Heart data

training sample 80, test sample 187, original feature dimension 44			
data name (dimension)	Initial partitioniing	test error (ICA)	test error (PCA)
Heart (44→2)	kd-Tree	18.18%	21.93%
Heart (44→2)	PCA-Tree	18.72%	19.25%
Heart (44→5)	kd-Tree	11.76%	16.04%
Heart (44→5)	PCA-Tree	17.11%	12.83%

Table 3. Comparison of accuracy on Iono data

training sample 200, test sample 150, original feature dimension 33			
data name (dimension)	Initial partitioning	test error (ICA)	test error (PCA)
Iono (33→5)	kd-Tree	6.00%	6.67%
Iono (33→5)	PCA-Tree	6.67%	5.33%
Iono (33→10)	kd-Tree	2.00%	2.67%
Iono (33→10)	PCA-Tree	2.00%	3.33%

Table 4. Comparison of accuracy on Sonar data

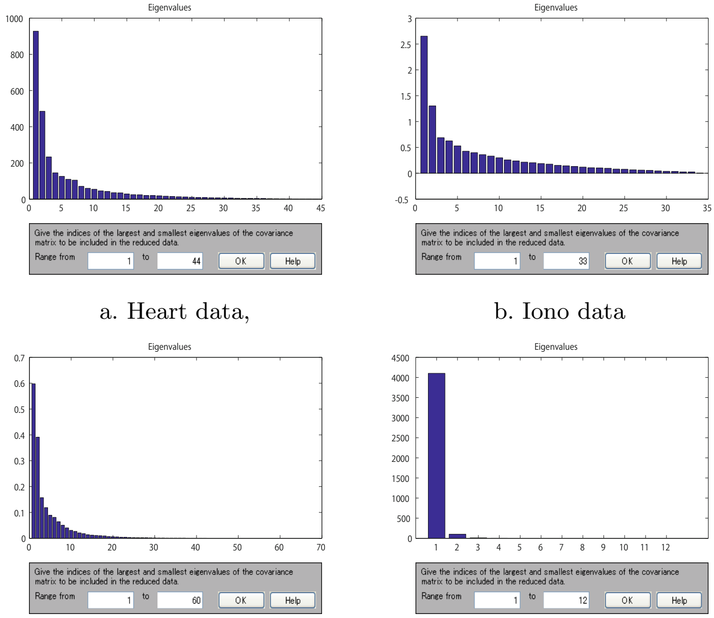
training sample 104, test sample 104, original feature dimension 60			
data name (dimension)	Initial partitioning	test error (ICA)	test error (PCA)
Sonar (60→2)	kd-Tree	36.54%	38.46%
Sonar (60→2)	PCA-Tree	40.38%	41.35%
Sonar (60→5)	kd-Tree	20.19%	28.85%
Sonar (60→5)	PCA-Tree	25.00%	25.96%

4.2 Wine Data

Table 5 shows the results which the test error rates of ICA were worse than those of PCA. Here, we discuss the characteristics of PCA and ICA-based dimension reduction in the data sets. Figure 1 shows the contribution ratio of eigenvalues in eigen decomposition. For example, the first five components account for 55.2% on Sonar data (Fig. 1-c), and the first two components account for 99.6% on Wine data (Fig. 1-d). From Fig. 1-d, the contribution ratio in the first principal component of Wine data was the majority compared to the other data (Herat, Iono, and Sona). These features imply that the Wine data can be summarized

Table 5. Comparison of accuracy on Wine data

training sample 2000, test sample 4497, original feature dimension 12			
data name (dimension)	Initial partitioning	test error (ICA)	test error (PCA)
Wine (12→5)	kd-Tree	5.56%	4.85%
Wine (12→5)	PCA-Tree	5.58%	4.76%
Wine (12→10)	kd-Tree	2.60%	1.58%
Wine (12→10)	PCA-Tree	1.65%	1.49%



Give the indices of the largest and smallest eigenvalues of the covariance matrix to be included in the reduced data.

a. Heart data,

Give the indices of the largest and smallest eigenvalues of the covariance matrix to be included in the reduced data.

b. Iono data

Give the indices of the largest and smallest eigenvalues of the covariance matrix to be included in the reduced data.

c. Sonar data,

Give the indices of the largest and smallest eigenvalues of the covariance matrix to be included in the reduced data.

d. Wine data

Fig. 1. Contribution ratio of eigenvalues

only by the first principal component and has no need to reformulate the ICA-based features. On the other hand, when the multi-dimensional data sets cannot be summarized by a solo feature value as in the case of the previous subsection, PCA scores should be further processed into ICA scores for achieving higher recognition rates.

These results fairly demonstrate the applicability of ICA-based preprocessing in FCMC.

5 Conclusion

In this paper, the applicability of ICA-based preprocessing to FCM classifier was investigated. The recognition rate of FCMC was compared using four benchmark data sets available from the UCI benchmark repository. As the result, in the most cases, the performances of ICA-based approach were better than those of PCA-based one, i.e., PCA-based preprocessed data sets should be further processed by ICA before applying FCMC when the data set can be summarized by two or more dimensional feature values.

Potential future work includes application of the ICA-based preprocessing model to much higher dimensional data sets [18].

Acknowledgment. This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under Grant-in-Aid for Scientific Research (23500283, 23500284).

References

1. Kobayashi, T., Ichihashi, H., Honda, K., Notsu, A.: Mixed Usage of MATLAB and Visual C for Improving Classification Time and Training Time of FCM Classifier. In: 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on Advanced Intelligent Systems, pp. 1994–1998 (2012)
2. Karhunen, J., Oja, E., Wang, L., Vigario, R., Joutsensalo, J.: A Class of Neural Networks for Independent Component Analysis. IEEE Trans. Neural Networks 8, 486–504 (1997)
3. Sproull, R.F.: Refinements to Nearest-neighbor Searching in k-dimensional Trees. Algorithmica 6, 579–589 (1991)
4. Friedman, J.H., Bentley, J.L., Finkel, R.A.: An Algorithm for Finding the Best Matches in Logarithmic Expected Time. ACM Trans. Math, Software 3(3), 209–226 (1977)
5. Bentley, J.L.: Multidimensional Binary Search Trees Used for Associative Searching. Communications of the ACM 18(9), 509–517 (1975)
6. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press (1981)
7. Miyamoto, S., Yasukochi, T., Inokuchi, R.: A Family of Fuzzy and Defuzzified c-Means Algorithms. In: International Conference on Computational Intelligence for Modelling, Control and Automation, pp. 170–176 (2005)

8. Miyamoto, S., Ichihashi, H., Honda, K.: Algorithms for Fuzzy Clustering, Methods in *c*-Means Clustering with Applications. Springer, Berlin (2008)
9. Ichihashi, H., Miyagishi, K., Honda, K.: Fuzzy *c*-Means Clustering with Regularization by K-L Information. In: 10th IEEE International Conference on Fuzzy Systems, vol. 3, pp. 924–927 (2001)
10. Tipping, M.E., Bishop, C.M.: Mixtures of Probabilistic Principal Component Analyzers. Neural Computation 11, 443–482 (1999)
11. Sun, F., Omachi, S., Aso, H.: Precise Selection of Candidates for Hand Written Character Recognition. IEICE Trans. Information and Systems E79-D(3), 510–515 (1996)
12. Ichihashi, H., Notsu, A., Honda, K.: Semi-hard *c*-Means Clustering with Application to Classifier Design. In: 2000 IEEE International Conference on Fuzzy Systems, pp. 2788–2795 (2010)
13. Ichihashi, H., Honda, K., Notsu, A.: Comparison of Scaling Behavior Between Fuzzy *c*-Means Based Classifier with Many Parameters and LibSVM. In: 2011 IEEE International Conference on Fuzzy Systems, pp. 386–393 (2011)
14. Hyvärinen, A., Oja, E.: A Fast Fixed-point Algorithm for Independent Component Analysis. Advances in Neural Information Processing Systems 9, 1483–1492 (1997)
15. Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M., Goodenday, L.S.: Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. Artificial Intelligence in Medicine 23(2), 149–169 (2001)
16. Sigillito, V.G., Wing, S.P., Hutton, L.V., Baker, K.B.: Classification of Radar Returns from the Ionosphere Using Neural Networks. Johns Hopkins APL Technical Digest 10, 262–266 (1989)
17. Gorman, R.P., Sejnowski, T.J.: Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets. Neural Networks 1, 75–89 (1988)
18. Ichihashi, H., Notsu, A., Honda, K., Miyamoto, E.: FCM Classifier for High-dimensional Data. In: 2008 IEEE International Conference on Fuzzy System, pp. 200–206 (2008)

Random Segmentation Based Principal Component Analysis to Remove Residual MR Gradient Artifact in the Simultaneous EEG/fMRI: A Preliminary Study

Hyun-Chul Kim and Jong-Hwan Lee*

Department of Brain Cognitive Engineering, Korea University
Seoul 136-713, Republic of Korea
{hyunchul_kim,jonghwan_lee}@korea.ac.kr

Abstract. In the electroencephalography (EEG) data simultaneously acquired with the functional magnetic resonance imaging (fMRI) data, the removal of the residual magnetic resonance (MR) gradient artifacts has been a challenging issue. To remove gradient artifacts generated from switching MR gradient field, average artifact subtraction (AAS) has been widely used. After applying the AAS method, however, residual MR gradient artifacts still remained in corrected EEG data. In this study, we proposed a novel method to remove the residual MR gradient artifacts (GAs) using random segmentation based principal component analysis (rsPCA). The performance of rsPCA was compared to that of the independent component analysis (ICA) method using data acquired from a motor imagery task. The results indicated that rsPCA could suppress further the residual MR gradient artifacts remained from the AAS step compared to the ICA method.

Keywords: Simultaneous EEG/fMRI, random segmentation, principal component analysis, electroencephalography, functional magnetic resonance imaging, MR gradient artifact.

1 Introduction

The simultaneous electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) technique has shown a great promise for investigating human brain function fulfilling both the superior temporal and spatial resolution [1]. However, it has long been an issue that the EEG data are corrupted by the artifacts induced from MR gradient switching during concurrent fMRI data acquisition [2]. In detail, using an gradient-echo echo planar imaging (EPI) pulse sequence to acquire fMRI data, MR gradient changes are repeatedly changing over the course of fMRI data acquisition, which results in repeated artifactual patterns in the simultaneously acquired EEG data. The amplitudes of these MR gradient artifacts (GAs) in the EEG data are several times greater than the EEG signal amplitude induced from a neuronal activity [2]. Thus, it is crucial to remove these MR-GAs from EEG data and to obtain meaningful EEG features associated with neuronal activity.

* Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 101–107, 2013.

An average artifact subtraction (AAS) method [3, 4] has been widely used for removal of these MR-GAs. The method utilizes the repeated patterns of the MR-GAs added in the EEG data to estimate an average MR-GA template. The estimated MR-GA template is then subtracted from the contaminated EEG data and consequently the substantial gradient-related artifacts were successfully attenuated using this approach. However, the residual MR-GAs [5] are still remained in EEG data due to the alternating slice timings across the fMRI volumes and mismatch in the hardware clocks between the EEG and fMRI systems. To further remove these residual MR-GAs, an independent component analysis (ICA) method in the context of a blind-signal separation (BSS-ICA) approach has been reported [6, 7]. In the resulting separating independent sources, the sources related to the MR-GAs were identified via a visual inspection and subsequently these sources were removed. However, the BSS-ICA approach has been limited to separate the purely residual MR-GAs in the separated ICs as often neuronal components were mixed in the separated MR-GA related IC [7].

To address this issue, we proposed a novel data-driven approach of random segmentation based principal component analysis (rsPCA), which is a standard PCA approach using data sets randomly segmented across a time-series data in each of the EEG channel to extract feature sets characterized the MR-GA. We hypothesized that our proposed rsPCA approach can extract the MR-GA related features determined based on a prior knowledge of frequency information of MR gradient switching. Also, we hypothesized that the rsPCA based MG-GA removal method can minimize potential signal loss. The resulting performance will be explicitly compared with that of the ICA approach in the context of the suppressing the MR-GA related spectral powers while maintaining the spectral powers potentially representing neuronal activity.

2 Materials and Methods

2.1 Data Acquisition

Four healthy right-handed volunteers participated in this study after providing written informed consents. The EEG data were acquired using a 32-channel MR-compatible EEG system (including a single electrocardiogram channel, BrainProducts GmbH, Germany) and blood oxygenation level dependent (BOLD) fMRI signals (TR/TE=1000/28ms, FOV=240×240mm2, matrix size=64×64, the number of interleaved slices=20 with no gap, thickness=7mm) while fMRI data were simultaneously acquired using a 12-channel head coil and gradient-echo echo-planar imaging (EPI) pulse sequence in a 3-T MRI (Tim Trio, Siemens, Erlangen, Germany). The EEG data were referenced to the FCz electrode and sampled at 5000Hz with a resolution of 0.5uV/bit. When the scanner was operated, 5V TTL pulse from the MRI scanner was sent to the EEG device, so the onset timings of each slice acquisition within a whole brain volume were recorded as makers on the EEG data that will be used for the AAS step.

2.2 Experimental Design

A block-based paradigm with right-hand motor imagery tasks was designed as a task paradigm. Four subjects completed an fMRI run (with a single fMRI scan) consisted of 10 trials of right-hand motor imagery tasks. As shown in Fig. 1, the fMRI run lasted 320 seconds and consisted of 10 task blocks (30s for each) followed by a rest block (20s). In the task block, there were two beeping alarm sounds during 0.5s to indicate the onset and offset of imagery task. When subjects heard the first beep (i.e., onset of the task), they were instructed to imagine their right-hand clenching movements at a pace of 3Hz for 2.5s until the last beeping (i.e., end of the task) was played. After the task, subjects were instructed to rest (26.5s) and they performed the right-hand motor imagery tasks in the remaining trials.

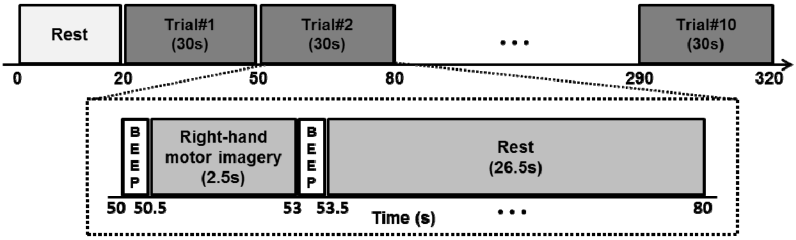


Fig. 1. Experiment design of the adopted right-hand motor imagery task

3 Random Segmentation Based PCA

Fig. 2 illustrates an overall flow diagram of our study. First, AAS method was applied to the MR-GA contaminated EEG data using Bergen EEG-fMRI toolbox (http://fmri.uib.no). In this study, a plugin developed for EEGLAB toolbox (http://sccn.ucsd.edu/ eeglabinto) was used. The resulting EEG data were down-sampled at 80Hz. Then, the rsPCA is applied to the down-sampled AAS applied EEG data.

As shown in the middle plot of the Fig. 2, the entire time-series from a single EEG channel was pseudo-randomly segmented. In this study, a time frame with 160 sample points (2s) was used as the length of each EEG segment. The duration of the EEG segment is two times longer than the duration of the fMRI volume acquisition (i.e. 1s) so it could capture the repeated patterns of the MR-GA across two fMRI volumes. A 2-D matrix (time-by-segments) with randomly segmented EEG data was then generated based on the concatenation of these randomly selected EEG segments. Using this 2-D data matrix, a PCA was conducted and eigenvalues and corresponding eigenvectors were estimated from a covariance matrix of the 2-D data matrix. Then, the estimated eigenvectors were transformed into the frequency domain using fast Fourier transformation (FFT) method with a window size of 160 to calculate the frequency spectrum. To select MR-GA related eigenvectors, frequency of the fMRI slice timing acquisitions were utilized. More specifically, the frequencies associated with the MR-GA are 20, 21.5Hz, and 40Hz (i.e. harmonic frequency of 20Hz; sample frequency=80Hz) since the periods of the slice acquisition were 50ms or 47.5ms. Thus, any eigenvectors (i.e. MR-GA features) whose center-frequencies are proximal (±10%) to the frequencies associated with the MR-GA, these eigenvectors were selected as potential MR-GA related features and subsequently removed from the corresponding channel of the EEG data. In detail, the coefficients of the eigenvectors were estimated from the EEG data of the corresponding channel via least-squares algorithm by minimizing the reconstruction error between (1) the original EEG data and (2) reconstructed EEG data which is a linear combination of the eigenvectors and corresponding coefficients. To remove the MR-GAs, the coefficients of the MR-GA related eigenvectors were set to zero and the EEG data were reconstructed. This procedure was independently applied to the EEG data in each channel and for each subject.

An Infomax based ICA algorithm implemented in the EEGLAB was also applied for performance evaluation. More specifically, a total of 32 ICs were extracted using the EEG data across the 32-channels and MR-GA related ICs were selected and subsequently excluded in the reconstruction of EEG data [6]. This ICA based MR-GA removal process is identically applied to the data sets from each subject.

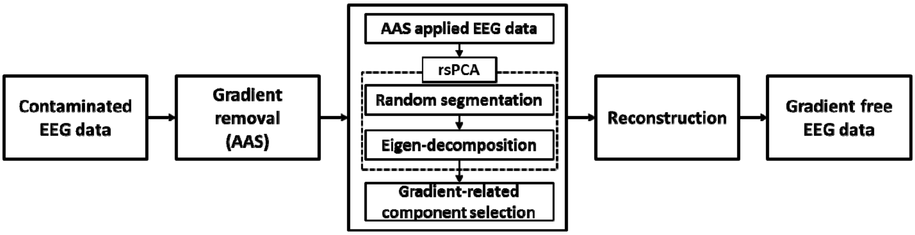


Fig. 2. The overall process to remove gradient artifact

4 Results

4.1 MR-GA Related Features from rsPCA

Fig. 3 represents that temporal patterns on Fz, Cz, and Pz channels before/after rsPCA process from a subject and two representative temporal/frequency patterns related with MR-GA. In rsPCA, 160 eigenvectors were estimated and the estimated eigenvectors were transformed into the frequency domain to determine MR-GA related features. As shown in the right panel of the Fig. 3, two representative MR-GA related eigenvectors were found on temporal/frequency domain among 160 eigenvectors. Using the MR-GA related frequencies (i.e., 20, 21.5Hz and 40Hz) estimated from the periods of the slice acquisition (i.e., 50ms and 47.5ms), several eigenvectors were selected. The selected eigenvectors were excluded during the reconstruction of EEG data. After the reconstruction, the rsPCA method showed that MR-GAs left by AAS were successfully removed.

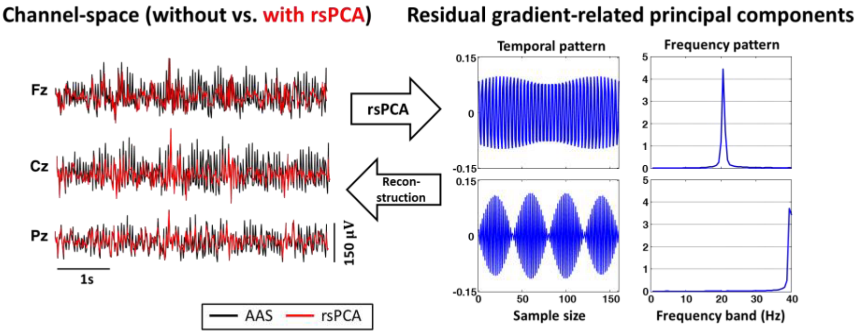


Fig. 3. Example results of residual MR-GA removal using an rsPCA approach

4.2 Spectral Profiles

Fig. 4 shows an example of EEG power spectrums from C3 channel after AAS, ICA, and rsPCA process. AAS method was applied into contaminated EEG signals across subjects. After applying the AAS method channel, substantial residual MR-GAs were removed by ICA and rsPCA. Subsequently, corrected EEG signals from each method were z-score normalized and transformed into the frequency domain to qualify evaluation across subjects.

As shown Fig 4, ICA result presented that frequency powers above 19Hz were relatively decreased compared to the rsPCA. Meanwhile, rsPCA showed that frequency powers were considerable overlaps with the frequency power of AAS result except the MR-GA frequency ranges (20, 21.5 or 40Hz). These results demonstrated that rsPCA achieved the effective artifact suppression with respect to the frequency power of the residual MR-GAs.

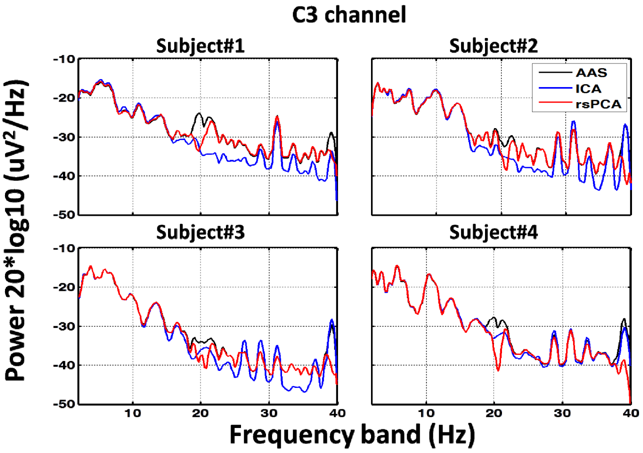


Fig. 4. The exemplified comparison result from ICA, rsPCA, and AAS applied EEG data

5 Discussion

In this study, we proposed a novel data-driven method to remove the residual MR-GAs using rsPCA. The performance of rsPCA was compared to that of ICA method with AAS applied EEG data sets. The performance results showed the spectral powers related MR-GAs were successfully suppressed by rsPCA while the spectral powers representing potential neuronal activity maintained. On the other hand, the result from the ICA method showed the attenuation of MR-GAs powers as well as non-artifact related signal powers. This meant that separated ICs were potentially mixed neuronal components. Consequently, rsPCA can effectively suppress the residual MR-GAs with minimization of potential signal loss compared to the ICA method.

However, there may be a potential issue of overestimation. As shown in Fig.4, a sharp decline pattern of spectral power at 20Hz was found in subject#4. That may be because a small time frame (160 sample points) was adopted to form a 2-D matrix for carrying out PCA. For example, if the time frame size is increased, eigenvectors estimated from PCA is more likely to finely separate spectral powers. These separated spectral powers would avoid from a sharp decline of spectral power during the feature selection of rsPCA. In addition, only use of temporal/frequency information may be unobvious which eigenvectors were highly related with MR-GAs due to spectral powers around the MR-GA related frequency potentially contain neuronal activity signals. To overcome this problem, statistical metrics and methods [10, 11] between estimated MR-GA from AAS and reconstructed signals from each eigenvector may be useful for feature selection.

Further work is warranted to utilize statistical metrics and methods (e.g., mutual information analysis, correlation analysis) and to change the number of sample size to avoid overestimation during the feature selection. Moreover, quantitative evaluation will be conducted by comparing AAS, ICA, and rsPCA method with a number of subjects to prove an efficacy of our method.

6 Conclusion

In this study, we proposed random segmentation based principal component analysis to effectively suppress residual artifacts left by AAS. The performance on power spectrum analysis showed that rsPCA method to successfully remove residual artifacts with severe signal loss compared to the ICA method. Our proposed rsPCA seems a promising to be able to contribute to enhance the quality of EEG data acquired during concurrent fMRI scanning.

Acknowledgments. This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (MEST) (2012-0002342) and in part by Samsung Electronics, Inc.

References

1. Horwitz, B., Poeppel, D.: How Can EEG/MEG and fMRI/PET Data Be Combined? Hum. Brain. Mapp. 17, 1–3 (2002)
2. Debener, S., Mullinger, K.J., Niazy, R.K., Bowtell, R.W.: Properties of the Ballistocardiogram Artefact as Revealed by EEG Recordings at 1.5, 3 and 7 T Static Magnetic Field Strength. Int. J. Psychophysiol. 67, 189–199 (2008)
3. Allen, P.J., Polizzi, G., Krakow, K., Fish, D.R., Lemieux, L.: Identification of EEG Events in the MR Scanner: The Problem of Pulse Artifact and a Method for Its Subtraction. NeuroImage 8, 229–239 (1998)
4. Allen, P.J., Josephs, O., Turner, R.: A Method for Removing Imaging Artifact from Continuous EEG Recorded during Functional MRI. NeuroImage 12, 230–239 (2000)
5. Niazy, R.K., Beckmann, C.F., Iannetti, G.D., Brady, J.M., Smith, S.M.: Removal of FMRI environment artifacts from EEG data using optimal basis sets. NeuroImage 28, 720–737 (2005)
6. Mantini, D., Perrucci, M.G., Cugini, S., Ferretti, A., Romani, G.L., Del Gratta, C.: Complete Artifact Removal for EEG Recorded during Continuous fMRI using Independent Component Analysis. NeuroImage 34, 598–607 (2007)
7. Lee, J.H., Oh, S., Jolesz, F.A., Park, H.W., Yoo, S.S.: Application of Independent Component Analysis for the Data Mining of Simultaneous EEG-fMRI: Preliminary Experience on Sleep Onset. Int. J. Neurosci. 119, 1118–1136 (2009)
8. Gonçalves, S.I., Pouwels, P.J.W., Kuijer, J.P.A., Heethaar, R.M., de Munck, J.C.: Artifact Removal in Co-registered EEG/fMRI by Selective Average Subtraction. Clin. Neurophysiol. 118, 2437–2450 (2007)
9. Garreffa, G., Carnì, M., Gualniera, G., Ricci, G.B., Bozzao, L., De Carli, D., Morasso, P., Pantano, P., Colonnese, C., Roma, V., Maraviglia, B.: Real-Time MR Artifacts Filtering during Continuous EEG/fMRI Acquisition. Magn. Reson. Imaging 21, 1175–1189 (2003)
10. Negishi, M., Abildgaard, M., Nixon, T., Constable, R.T.: Removal of Time-Varying Gradient Artifacts from EEG Data Acquired during Continuous fMRI. Clinical Neurophysiol. 115, 2181–2192 (2004)
11. Liu, Z., de Zwart, J.A., van Gelderen, P., Kuo, L.W., Duyn, J.H.: Statistical Feature Extraction for Artifact Removal from Concurrent fMRI-EEG Recordings. NeuroImage 59, 2073–2087 (2012)

Extracting Latent Dynamics from Multi-dimensional Data by Probabilistic Slow Feature Analysis

Toshiaki Omori

Department of Electrical and Electronic Engineering,
Graduate School of Engineering, Kobe University
1–1, Rokkodai-cho, Nada-ku, Kobe 657–8501 Japan
omori@eedept.kobe-u.ac.jp
http://www2.kobe-u.ac.jp/~omoritos/

Abstract. Slow feature analysis (SFA) is a time-series analysis method for extracting slowly-varying latent features from multi-dimensional data. In this paper, the probabilistic version of SFA algorithms is discussed from a theoretical point of view. First, the fundamental notions of SFA algorithms are reviewed in order to show the mechanism of extracting the slowly-varying latent features by means of the SFA. Second, recent advances in the SFA algorithms are described on the emphasis of the probabilistic version of the SFA. Third, the probabilistic SFA with rigorously derived likelihood function is derived by means of belief propagation. Using the rigorously derived likelihood function, we simultaneously extracts slow features and underlying parameters for the latent dynamics. Finally, we show using synthetic data that the probabilistic SFA with rigorously derived likelihood function can estimate the slow feature accurately even under noisy environments.

Keywords: Slow feature analysis, State-space model, Probabilistic information processing, Bayesian statistics, Latent dynamics.

1 Introduction

Slow feature analysis (SFA) is a time-series analysis method for extracting slowly varying features from multi-dimensional data [1]. In recent years, the SFA has attracted much attention in computational neuroscience studies to establish models for complex cells in the visual systems, and those for place cells and grid cells in the hippocampus and entorhinal cortex [2–4]. In those models, the SFA has been used under assumption that slowly varying features play an important role in the information processings in our brain. Moreover, the SFA has been applied to important machine learning problems such as pattern recognition and feature extraction from high-dimensional data and so on [5–8].

Recently, a probabilistic version of the SFA has been proposed using the framework of state-space model [9]; probabilistic perspective has been useful for many

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 108–116, 2013.

machine learning algorithms such as principal component analysis [10], independent component analysis [11] and so on. In the conventional probabilistic SFA [9], a likelihood function used to estimate parameters of the model is approximately evaluated by assuming that there exists no observation noise. Although observed data that we have to deal with would be noisy in general, it has been unclear whether the slow feature estimated by the conventional probabilistic SFA is accurate for such noisy data. Actually, a recent theoretical study showed that the conventional method cannot extract slow features accurately under noisy environments [13].

In this paper, we discuss recent advances in the SFA algorithms from a theoretical point of view. First, the fundamental notions of SFA algorithms are reviewed from a theoretical point of view in order to show how the SFA extracts the slowly-varying latent features. Second, recent advancements in the SFA algorithms are discussed on the emphasis of the probabilistic version of the SFA. The probabilistic SFA with rigorously derived likelihood function is derived by using belief propagation [12]; the belief propagation is a method to realize rigorous results for the graphical model with no loops while we can find the probabilistic SFA has no loops in its graphical structure. Using the rigorously derived likelihood function, a probabilistic version of SFA considering the effect of observation noise is realized and we can simultaneously extract slow features and underlying parameters for the latent dynamics. Finally, we show using synthetic data that the probabilistic SFA with rigorously derived likelihood function can estimate the slow feature accurately even under noisy environments.

2 Theory

In this section, we first review conventional SFA algorithms and then discuss recent advances in the SFA algorithm employing a probabilistic framework with rigorously derived likelihood function by means of belief propagation. We show that the probabilistic SFA with rigorously derived likelihood function realizes accurate and robust estimation of the slow feature and parameters even under noisy environments.

2.1 Deterministic SFA

The original SFA is a deterministic algorithm to extract the most slowly varying feature (called "slow feature") from multi-dimensional time series data [1]. A schematic diagram of the deterministic SFA is shown in Fig. 1 (a).

For multi-dimensional input time series data $\boldsymbol{x}(t) \in \mathbb{R}^M$, the output of the SFA $y_j(t)$ $(j = 1, \cdots, N)$ is obtained using transformation $y_j(t) = g_j(\boldsymbol{x}(t))$. This transformation $g_j(\boldsymbol{x})$ is determined to minimize the following expression:

$$\Delta(y_j) = \langle \dot{y}_j^2 \rangle_t \tag{1}$$

where $\Delta(y_j)$ is called Δ-value and $\langle \cdot \rangle_t$ denotes an average with respect to time. Namely, in the deterministic SFA, we perform transformation $g_j(\boldsymbol{x})$ which minimizes the derivative of output $\boldsymbol{y}(t)$ with respect to time t. Within outputs of

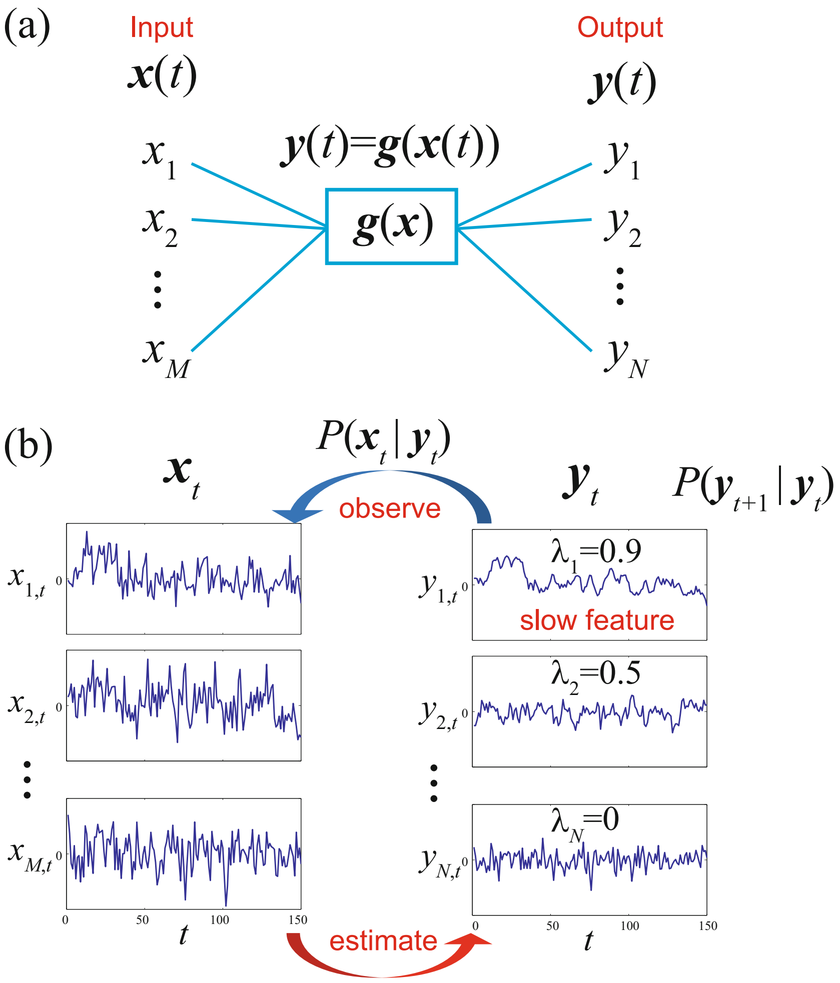


Fig. 1. A schematic diagram of slow feature analysis (SFA). (a) In the deterministic SFA, multi-dimensional time series data $\boldsymbol{x}(t)$ are transformed via scalar functions $g_j(\boldsymbol{x})$ into outputs $y_j(t)$. An element of outputs with minimal $\varDelta$-value corresponds to slow feature. (b) In the probabilistic SFA, latent variables $\boldsymbol{y}_t$ are estimated from observed variable $\boldsymbol{x}_t$ by using framework of Bayesian statistics. The latent variable $\boldsymbol{y}_{i,t}$ with the largest λ_i corresponds to the slow feature.

the SFA $\{y_j(t)\}$, the output $y_j(t)$ with minimum Δ-value corresponds to a slow feature.

Additionally, the constraint conditions are employed:

$$\langle y_j \rangle_t = 0, \quad \text{(zero mean)} \tag{2}$$

$$\langle y_j^2 \rangle_t = 1, \quad \text{(unit variance)} \tag{3}$$

$$\langle y_i y_j \rangle_t = 0, \quad \text{(decorrelation)} \tag{4}$$

These three constraints are employed to normalize output signals, avoid trivial results, and guarantees that different output elements have different information.

2.2 Probabilistic SFA

Recently, a probabilistic version of SFA has been proposed [9]. However, the likelihood function is approximately derived by assuming no observation noise in the conventional probabilistic SFA [9]. Actually, a theoretical study showed that the conventional SFA cannot estimate the slow feature accurately under noisy environment [13].

Here we discuss a probabilistic SFA with rigorous derivation of likelihood function [14, 15]. We consider extraction of N-dimensional latent variables $\boldsymbol{y}_t$ from M-dimensional observed variables $\boldsymbol{x}_t$ based on probability distribution reflecting the deterministic SFA (Fig. 1(b)).

Latent variables $\boldsymbol{y}_t$ including the slow feature is described by system model:

$$\boldsymbol{y}_t = \boldsymbol{\lambda} \boldsymbol{y}_{t-1} + \boldsymbol{\eta}_t. \tag{5}$$

Namely, the latent variable $\boldsymbol{y}_t$ at each time depends on that $\boldsymbol{y}_{t-1}$ at the preceding time. $\boldsymbol{\lambda}$ is a parameter for the degree of the dependence of latent variable, and is expressed by a diagonal matrix with elements λ_n for the corresponding latent variables $y_{n,t}$. $\boldsymbol{\eta}_t$ describes a system noise obeying white Gaussian noise with average $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a diagonal matrix with elements $\sigma^2_{1:N}$. Thus, the dynamics of latent variables has two kinds of parameters: $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$. Here each element λ_n of $\boldsymbol{\lambda}$ takes a value between 0 and 1. Note that the dynamics of $y_{n,t}$ is slow for large value of λ_n whereas the dynamics of $y_{n,t}$ is fast for small value of λ_n. Therefore, the latent variable with the largest value of λ_n corresponds to slow feature to be extracted in the probabilistic SFA.

Observed variables $\boldsymbol{x}_t$ are assumed to be expressed using observation model:

$$\boldsymbol{x}_t = \boldsymbol{W}^{-1} \boldsymbol{y}_t + \boldsymbol{\varepsilon}_t \tag{6}$$

Namely, observed variables $\boldsymbol{x}_t$ are generated from the latent variables $\boldsymbol{y}_t$ converted by $M \times N$ matrix $\boldsymbol{W}^{-1}$ under observation noise $\boldsymbol{\varepsilon}_t$. The observation noise is assumed to be white Gaussian noise with average $\mathbf{0}$ and covariance $\sigma_x^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix.

In the probabilistic SFA, the latent variables $\boldsymbol{y}_t$ and the observation variables $\boldsymbol{x}_t$ are described by the state space model consisting of system model (Eq. (5))

and observation model (Eq. (6)). The state space model of the probabilistic SFA can be expressed by using probability density functions as follows:

$$p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{\lambda}\boldsymbol{y}_{t-1}, \boldsymbol{\Sigma}) \tag{7}$$

$$p(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{W}, \sigma_x^2\boldsymbol{I}) = \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{W}^{-1}\boldsymbol{y}_t, \sigma_x^2\boldsymbol{I}) \tag{8}$$

2.3 Rigorous Derivation of Likelihood Function by Belief Propagation

To estimate the latent variables $\boldsymbol{y}_t$, we need to estimate parameters in the state state model: $\boldsymbol{\theta} = \{\boldsymbol{W}^{-1}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \sigma_x^2\}$. For this purpose, the likelihood function of the probabilistic SFA is derived by means of the belief propagation [12, 14, 15].

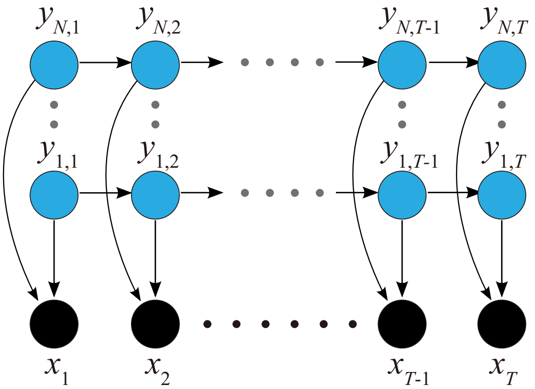


Fig. 2. A graphical structure of the probabilistic SFA. Each observed variable $\boldsymbol{x}_t$ at time t depends on a latent variable $\boldsymbol{y}_t$ at the same time t, whereas the latent variable $\boldsymbol{y}_t$ at time t depends on the latent variable $\boldsymbol{y}_{t-1}$ at the preceding time $t-1$. The graphical structure of the probabilistic SFA has a straight structure and no loops. Based on this graphical structure, we perform belief propagation to derive the likelihood function.

The likelihood function of the probabilistic SFA obeys the following expression:

$$p(\boldsymbol{x}_{1:T}|\boldsymbol{\theta}) = \int d\boldsymbol{y}_{1:T} \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{W}, \sigma_x^2\boldsymbol{I}) p(\boldsymbol{y_1}|\boldsymbol{P}_0) \prod_{t=2}^{T} p(\boldsymbol{y_t}|\boldsymbol{y_{t-1}}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}) \tag{9}$$

To evaluate this likelihood function., Turner and Sahani [9] employed an approximation by assuming that observation noise σ_x^2 is zero. In the approximation, the probabilistic model of observation model (Eq. (8)) becomes Dirac's delta function, and integration in the likelihood function can be easily performed. However, this approximation assumes that there exists no observation noise, and estimation accuracy would be lowered for noisy data [13].

Here we discuss the rigorous derivation of the likelihood function in the probabilistic SFA by means of belief propagation [12, 13, 15]. Eq. (9) includes high-dimensional integration of joint distribution with respect to $\boldsymbol{y}_t$. As shown in Fig. 2, the graphical model of the probabilistic SFA has a straight structure and no loops. Based on this structure, we overcome the difficulty in high-dimensional integration by the belief propagation. Thus we can perform integration with respect to each latent variable $\boldsymbol{y}_t$ per time subsequently from time $t = 1$. Since $\boldsymbol{y}_t$ depends on $\boldsymbol{y}_{t-1}$, integrations after time $t = 2$ can be performed by using the integration for the preceding time. By the belief propagation, a marginal distribution $\alpha(\boldsymbol{y}_t)$ can be propagated as a message from time $t = 1$ to time $t = T$ as the following recursion relation:

$$c_t \alpha(\boldsymbol{y}_t) = p(\boldsymbol{x}_t|\boldsymbol{y}_t) \int d\boldsymbol{y}_{t-1} \alpha(\boldsymbol{y}_{t-1}) p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) \tag{10}$$

Since probability distributions in the above expression obey Gaussian distributions, the marginal distribution $\alpha(\boldsymbol{y}_t)$ becomes a Gaussian distribution,

$$\alpha(\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_t|\boldsymbol{\mu}_t, \boldsymbol{V}_t) \tag{11}$$

Here coefficient c_t is conditional distribution of observation model as follows:

$$c_t = p(\boldsymbol{x}_t|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}) \tag{12}$$

By conducting analytical treatments using the belief propagation, we rigorously derive the likelihood of the probabilistic SFA as follows:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{W}^{-1}\boldsymbol{\lambda}\boldsymbol{\mu}_{t-1}, \boldsymbol{Z}_{t-1}) \tag{13}$$

where

$$\begin{aligned} \boldsymbol{Z}_t &= \boldsymbol{W}^{-1}\boldsymbol{P}_t\boldsymbol{W}^{-1^T} + \sigma_x^2\boldsymbol{I}, & \boldsymbol{P}_t &= \boldsymbol{\Sigma} + \boldsymbol{\lambda}\boldsymbol{V}_t\boldsymbol{\lambda}^T \\ \boldsymbol{V}_t &= (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{W}^{-1})\boldsymbol{P}_{t-1}, & \boldsymbol{\mu}_t &= \boldsymbol{\lambda}\boldsymbol{\mu}_{t-1} + \boldsymbol{K}_t(\boldsymbol{x}_t - \boldsymbol{W}^{-1}\boldsymbol{\lambda}\boldsymbol{\mu}_{t-1}) \end{aligned} \tag{14}$$

Here $\boldsymbol{K}_t$ is a Kalman gain matrix and is shown to depend on observation noise σ_x. The rigorously derived likelihood function $p(\boldsymbol{x}|\boldsymbol{\theta})$ is a product of c_t and average and covariance of each c_t can be obtained from that of the preceding time. In this rigorous framework, we estimate parameters by using the derived likelihood function and obtain the latent variables $\boldsymbol{y}_t$ including slow feature. Note that the approximated likelihood function used in the probabilistic SFA proposed by Turner and Sahani [9] can be obtained in the limit of $\sigma_x \to 0$.

Using the rigorously derived likelihood function and the state space model of the SFA, the probabilistic SFA realizes the simultaneous estimation of the slow feature and its underlying parameters.

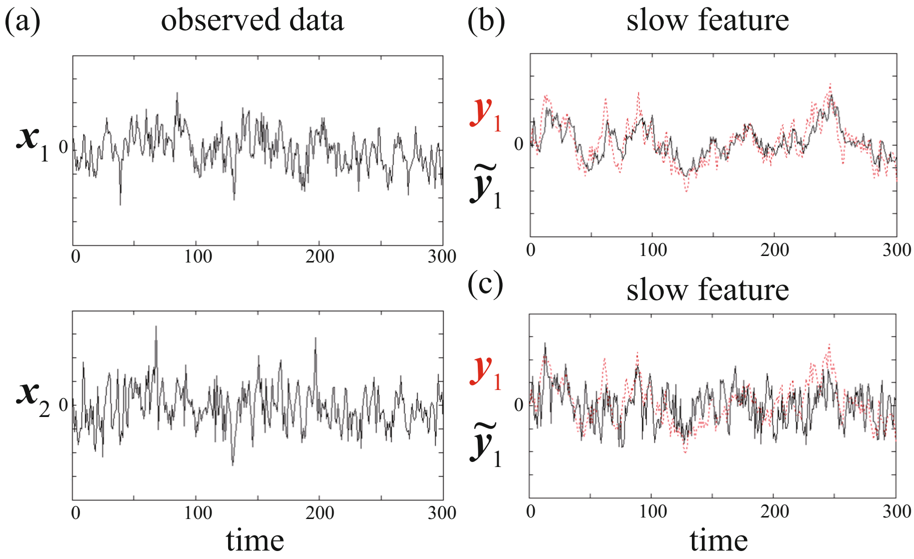


Fig. 3. Estimated results using probabilistic SFAs. (a) Parts of multi-dimensional observed data $\boldsymbol{x}_t$. (b) Estimated slow feature by using the probabilistic SFA with rigorously derived likelihood function $\tilde{\boldsymbol{y}}_1$ (black solid line) and true slow feature $\boldsymbol{y}_1$(red dotted line). (c) Estimated slow feature $\tilde{\boldsymbol{y}}_1$ (black solid line) by using the probabilistic SFA with approximated likelihood function and true slow feature $\boldsymbol{y}_1$(red dotted line).

3 Results

In this section, we compare the performance of the probabilistic SFAs. Both latent and observation variables are generated numerically based on the probabilistic SFA. The latent variables $\boldsymbol{y}_t$ and the parameters $\boldsymbol{\theta} = \{\boldsymbol{W}^{-1}, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \sigma_x^2\}$ are estimated using the probabilistic SFA with the likelihood function rigorously derived in the previous section. For simplicity, the dimension of observed variables $\boldsymbol{x}_t$ is set to be the same as that of latent variables $\boldsymbol{y}_t$.

3.1 Estimation of Slow Feature from Multi-dimensional Data

Here we extract a slow feature from noisy observed data by using the probabilistic SFA with rigorously derived likelihood function. The slow feature $\tilde{\boldsymbol{y}}_t$ is estimated from multi-dimensional time-series data $\boldsymbol{x}_t$ (Fig. 3 (a)). As shown in Fig. 3 (b), the estimated slow feature $\tilde{\boldsymbol{y}}_t$ (black solid line) exhibits similar dynamical behaviors shown in the true slow feature $\boldsymbol{y}_t$. In contrary, the slow feature estimated by the probabilistic SFA with approximated likelihood function is less similar to the true one (Fig. 3 (c)). From these results, we find that the probabilistic SFA with rigorously derived likelihood function [15] gives

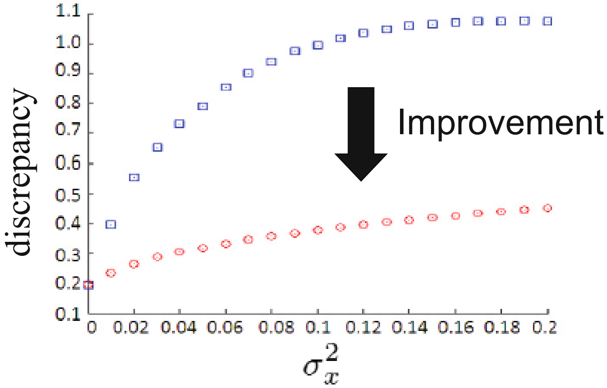


Fig. 4. Comparison between probabilistic SFA with rigorously derived likelihood function (circles) and conventional one with approximated likelihood function(squares). Discrepancy between the true slow feature $\boldsymbol{y}_1$ and the estimated slow feature $\tilde{\boldsymbol{y}}_1$ is evaluated for different values of observation noise σ_x^2. We find that the probabilistic SFA with rigorously derived likelihood function shows better performance than conventional one approximated likelihood function [9] for noisy data.

better performance compared with conventional one with approximated likelihood function [9].

3.2 Effect of Observation Noise on Performance

To evaluate the effect of observation noise on estimation performance, we perform estimation for different levels of observation noise. Figure 4 shows how the discrepancy between the estimated and the true slow feature changes depending on the observation noise. We find that the probabilistic SFA with rigorously derived likelihood function gives better performance than conventional one with approximated likelihood function; even though the results of two methods are similar when there exists no observation noise, estimation errors for the probabilistic SFA with rigorously derived likelihood function are much smaller than those for conventional one with approximated likelihood function. From these results, we find that the probabilistic SFA with rigorously derived likelihood function extracts slow features more accurately.

4 Concluding Remarks

In this paper, we discussed the latent dynamics extraction algorithms using the SFA framework. We first described the basic framework of the deterministic SFA and the probabilistic SFA, and then discussed the recent advances in the probabilistic SFA. The likelihood function of the probabilistic SFA has been derived rigorously by means of belief propagation, while the likelihood function was approximately evaluated by assuming that observation noise is zero in the conventional SFA algorithm. Furthermore, we have shown using numerical data that the probabilistic SFA with rigorously derived likelihood function can estimate the slow feature and its underlying parameters for latent dynamics accurately even under noisy environments.

Acknowledgments. The author is grateful to special session chairs of the ICONIP 2013, especially to Prof. Seiichi Ozawa, for the invitation of the present paper to the special session. The author thanks and Prof. Masato Okada and Tomoki Sekiguchi of the University of Tokyo for valuable discussion throughout this study. This work was partially supported by Grants-in-Aid for Scientific Research on Innovative Areas (Nos. 22120506 and 25120010) and for Young Scientists (No. 25730147) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Wiskott, L., Sejnowski, T.J.: Slow Feature Analysis: Unsupervised Learning of Invariances. Neural Comput. 14, 715–770 (2002)
2. Berkes, P., Wiskott, L.: Slow Feature Analysis Yields a Rich Repertoire of Complex Cell Properties. J. Vis. 5, 579–602 (2005)
3. Franzius, M., Sprekeler, H., Wiskott, L.: Slowness and Sparseness Lead to Place, Head-Direction, and Spatial-View Cells. PLoS Comput. Biol. 3, 1605–1622 (2007)
4. Sprekeler, H., Wiskott, L.: A Theory of Slow Feature Analysis for Transformation-Based Input Signals with an Application to Complex Cells. Neural Comput. 23, 303–335 (2011)
5. Berkes, P.: Pattern Recognition with Slow Feature Analysis. Cognitive Sciences EPrint Archive (CogPrint) 4104 (2005)
6. Franzius, M., Wilbert, N., Wiskott, L.: Invariant Object Recognition with Slow Feature Analysis. In: Kůrková, V., Neruda, R., Koutník, J. (eds.) ICANN 2008, Part I. LNCS, vol. 5163, pp. 961–970. Springer, Heidelberg (2008)
7. Legenstein, R., Wilbert, N., Wiskott, L.: Reinforcement Learning on Slow Features of High-Dimensional Input Streams. PLoS Computational Biology 6, 1–13 (2010)
8. Huang, Y.P., Zhao, J.L., Liu, Y.H., Luo, S.W., Zou, Q., Tian, M.: Nonlinear Dimensionality Reduction Using a Temporal Coherence Principle. Inform. Sci. 181, 3284–3307 (2011)
9. Turner, R., Sahani, M.: A Maximum-Likelihood Interpretation for Slow Feature Analysis. Neural Comput. 19, 1022–1038 (2007)
10. Tipping, M.E., Bishop, C.M.: Probabilistic Principal Component Analysis. J. Royal Stat. Soc. 61, 611–622 (1999)
11. Beckmann, C.F., Smith, S.M.: Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. IEEE Trans. Med. Im. 23, 137–152 (2004)
12. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufman (1988)
13. Sekiguchi, T., Omori, T., Okada, M.: Effect of Observation Noise in Probabilistic Slow Feature Analysis. IPSJ Trans. Math. Model. Appl. (in press)
14. Sekiguchi, T., Omori, T., Okada, M.: Belief Propagation for Probabilisitic Slow Feature Analysis. IEICE Tech. Rep. (2011)
15. Omori T., Sekiguchi, T., Okada, M.: (in prep.)

Challenges in Representation Learning: A Report on Three Machine Learning Contests

Ian J. Goodfellow[1], Dumitru Erhan[2], Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio

[1] Université de Montréal, Montréal QC H3T 1N8, Canada
goodfeli@iro.umontreal.ca
[2] Google, Venice, CA 90291, USA
dumitru@google.com

Abstract. The ICML 2013 Workshop on Challenges in Representation Learning[1] focused on three challenges: the black box learning challenge, the facial expression recognition challenge, and the multimodal learning challenge. We describe the datasets created for these challenges and summarize the results of the competitions. We provide suggestions for organizers of future challenges and some comments on what kind of knowledge can be gained from machine learning competitions.

Keywords: representation learning, competition, dataset.

1 Introduction

This paper describes three machine learning contests that were held as part of the ICML workshop "Challenges in Representation Learning." The purpose of the workshop, organized by Ian Goodfellow, Dumitru Erhan, and Yoshua Bengio, was to explore the latest developments in representation learning, with a special emphasis on testing the capabilities of current representation learning algorithms (See [1] for a recent review) and pushing the field towards new developments via these contests. Ben Hamner and Will Cukierski handled all issues related to Kaggle hosting and ensured that the contests ran smoothly. Ian Goodfellow and Dumitru Erhan provided baseline solutions to each challenge, mostly in Pylearn2 [2] format. Google provided prizes for all three contests. The winner of each contest received $350 while the runner-up received $150. A diverse range of competitors spanning academia, industry, and amateur machine learning provided excellent solutions to all three problems. In this paper, we summarize their solutions, and discuss what we can learn from them.

[1] http://deeplearning.net/icml2013-workshop-competition

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 117–124, 2013.

2 The Black Box Learning Challenge

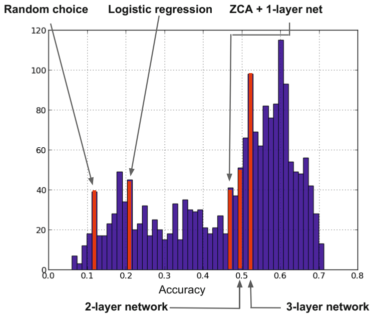


Fig. 1. Histogram of accuracies obtained by different submissions on the BBL-2013 dataset. Organizer-provided baselines shown in red.

The black box learning challenge[2] was designed with two goals in mind. First, the data was obfuscated, so that competitors could not use human-in-the-loop techniques like visualizing filters to guide algorithmic development. A common criticism of deep learning is that it is an art requiring an expert practitioner. By keeping the domain of the data secret, this contest reduced the usefulness of the human practitioner. This idea was similar to a recent DARPA-organized unsupervised and transfer learning challenge [3] which used obfuscated data and required submission of a representation of the data that would then be used on the competition server to train a very weak classifier. In this contest, we allowed competitors to use any method; using representation learning was not a requirement. The second goal of this contest was to test the ability of algorithms to benefit from extra unsupervised data. To this end, we provided only very few labeled examples.

This contest introduced the Black Box Learning 2013 (BBL-2013) dataset. The scripts needed to re-generate it are available for download[3]. The dataset is an obfuscated subset of the second (MNIST-like) format of the Street View House Numbers dataset[4]. Dumitru Erhan created the dataset. The original data contained 3,072 features (pixels) which he projected down to 1875 by multiplication by a random matrix. He also removed one class (the “4”s). These measures obfuscated the data so competitors did not know what task they were solving. The organizers did not reveal the source of the dataset until after the contest was over. To make the challenge emphasize semi-supervised learning, only 1,000 labeled examples were kept for training. Another 5,000 were used for the public leaderboard. For these examples, the labels are not provided to the competitors, but the features are. Each team may upload predictions for these examples twice per day. The resulting accuracy is published publicy. The public test set is thus a sort of validation set, but also gives one’s competitors information. Another 5,000 examples were used for the private test set. The features for these examples are given to the competitors as well, but only the contest administrators see the accuracy on them until after the contest has ended. The

[2] `http://www.kaggle.com/c/challenges-in-representation-learning-the-black-box-learning-challenge`

[3] `http://www-etud.iro.umontreal.ca/~goodfeli/bbl2013.html`

private test set is used to determine the winner of the contest. We also provided 130,000 unlabeled examples drawn from a set specified to be "less difficult" by the creators of SVHN.

218 teams submitted 1963 entries to the contest. 75 teams beat the best baseline (a 3-layer MLP) provided by the organizers. See Fig. 1 for a histogram of all the teams' performance. David Thaler won the contest with an accuracy of 70.22% using blending of three models that used sparse filtering[5] for feature learning, random forests for feature selection [6], and support vector machines[7] for classification. Other competitors such as Lukasz Romaszko [8] also obtained very competitive results with sparse filtering. This was an interesting outcome because sparse filtering has usually been perceived as an inexpensive and simple method that gives good but not optimal results. David Thaler and Lukasz Romaszko both observed that learning the sparse filtering features on the combination of the labeled and unlabeled data worked *worse* than learning the features on just the labeled data. This may be because the labeled data was drawn from the more difficult portion of the SVHN dataset. Dong-Hyun Lee [9] finished second in the contest, having independently rediscovered entropy regularization [10]. This very simple means of semi-supervised learning proved surprisingly effective and merits more attention. In third place, Dimitris Athanasakis and John Shawe-Taylor developed a new feature section / combination mechanism combined with MKL. Other top scorers included Jingjing Xie, Bing Xu and Zhang Chuang, who developed ensemble voting techniques for use with denoising autoencoders [11] and maxout networks [12].

A recent trend in deep learning has been to forego unsupervised learning entirely following recent improvements to discriminative training. This is probably a result of most datasets having several labeled examples. In this contest, with only 1,000 labeled training examples, most of the top scorers still needed to make use of the unlabeled data in some way.

3 The Facial Expression Recognition Challenge

In the facial expression recognition challenge[4] we invited competitors to design the best system for recognizing which emotion is being expressed in a photo of a human face. In this contest, we wanted to compare methods on a task that is well studied but using a completely new dataset. This avoids issues of overfitting to the test set of a repeatedly used benchmark dataset. One reason to hold such a contest is that it allows us to compare feature learning methods to hand-engineered features in as fair a manner as possible.

This contest introduced the Facial Expression Recognition 2013 (FER-2013) dataset. It is available for download[5]. FER-2013 was created by Pierre Luc Carrier and Aaron Courville. It is part of a larger ongoing project. The dataset was created using the Google image search API to search for images of faces

[4] `http://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge`

[5] `http://www-etud.iro.umontreal.ca/~goodfeli/fer2013.html`

that match a set of 184 emotion-related keywords like “blissful”, “enraged,” etc. These keywords were combined with words related to gender, age or ethnicity, to obtain nearly 600 strings which were used as facial image search queries. The first 1000 images returned for each query were kept for the next stage of processing. OpenCV [13] face recognition was used to obtain bounding boxes around each face in the collected images. Human labelers than rejected incorrectly labeled images, corrected the cropping if necessary, and filtered out some duplicate images. Approved, cropped images were then resized to 48x48 pixels and converted to grayscale. Mehdi Mirza and Ian Goodfellow prepared a subset of the images for this contest, and mapped the fine-grained emotion keywords into the same seven broad categories used in the Toronto Face Database [14]. The resulting dataset contains 35887 images, with 4953 “Anger” images, 547 “Disgust” images, 5121 “Fear” images, 8989 “Happiness” images, 6077 “Sadness” images, 4002 “Surprise” images, and 6198 “Neutral” images.

Ian Goodfellow performed some small-scale experiments to estimate the human performance on this task. He collected 1500 images of members of the LISA lab acting out the seven facial expressions. This dataset contains no label noise per se, though poor acting abilities mean that the Bayes rate could be quite high. On this dataset, human accuracy was 68±5%. FER-2013 could theoretical suffer from label errors due to the way it was collected, but Ian Goodfellow found that human accuracy on FER-2013 was 65±5%. While there may be label errors, they do not make the task significantly harder, at least not for a human. James Bergstra also determined the best performance of a “null” model, consisting of a convolutional network with no learning except in the final classifier layer. Using the TPE hyperparameter optimization algorithm, he found that the best such convolutional network obtains an accuracy of 60%. Using an ensemble of such models, he obtained an accuracy of 65.5%. See [15] for details.

56 teams submitted on the final dataset. Of these, four beat the best “null” ensemble model (which was not presented until after the contest was over–many more teams beat the simpler baselines provided by the organizers). Their scores are presented in Table 1. The top three teams all used convolutional neural networks [16] trained discriminatively with image transformations. The winner, Yichuan Tang, used the primal objective of an SVM as the loss function for training. This loss function has been applied to neural networks before, but he additionally used the L2-SVM loss function, a new development that gave great results on the contest dataset and others.

One of the questions we hoped to answer in this workshop is whether or not feature learning algorithms are ahead of other methods. Radu Ionescu, Marius Popescu, and Cristian Grozea provided the strongest submission that did not use feature learning. Their approach used SIFT [17] and MKL. This approach put their performance close to that of Maxim Milakov, who submitted the third best convolutional network. These results suggest that convolutional networks are indeed capable of outperforming hand-designed features, but the difference in accuracy is not extreme. It’s unclear whether the performance of the best deep network has reached the Bayes rate on this task or not.

Table 1. Private test set accuracy on FER-13

Team	Members	Accuracy
RBM [18]	Yichuan Tang	71.162%
Unsupervised	Yingbo Zhou, Chetan Ramaiah	69.267%
Maxim Milakov[6]	Maxim Milakov	68.821%
Radu + Marius + Cristi [19]	Radu Ionescu, Marius Popescu, Cristian Grozea	67.484%

4 The Multimodal Learning Challenge

The multimodal learning challenge[7] was intended to spur development of algorithms that discover a unified semantic representation of examples that have more than one input representation. In this case, the two input modalities were images and text.

Competitors were advised to use the small ESP game dataset [20] for training data, but all public sources of training data were allowed. The small ESP game dataset consists of 100,000 images of varying sizes that were annotated by players of an online game. Each image is tagged with on average 14 words, with a vocabulary of over 4,000 words.

In order to provide a new test set, Ian Goodfellow manually labeled 1,000 images obtained by Google image search queries for some of the most commonly used words in the small ESP game dataset. The labels were intended to resemble those in the training set. For example, they include incorrect spellings that were common in the training set. This dataset is available for download[8].

Kaggle does not yet provide the kinds of evaluation metrics typically used for multimodal learning, so the organizers devised a multimodal classification task. Each test image would be accompanied by two labels from the test set, with the classification task being to report which of the two labels is correct. Unfortunately, because this is a matching task, it proved too easy to yield interesting machine learning results. Yichuan Tang found that a base classifier with low accuracy could be coupled with the Hungarian algorithm to compute the optimal matching. The optimal matching constructed in this way obtained 100% accuracy. The contest ended in a three-way tie with 100% test accuracy. The winners were “RBM” (Yichuan Tang), “MMDL” [21] (Fangxiang Feng, Ruifan Li, and Xiaojie Wang), and “AlbinoSnowman” (John Park). RBM won the tie by submitting the first perfect solution. The tie between MMDL and AlbinoSnowman was broken because MMDL submitted a model file for verification and AlbinoSnowman did not. If a similar contest is organized in the future, we recommend labeling twice as many test images as are needed, then discarding half

[6] `http://nnforge.org`

[7] `http://www.kaggle.com/c/challenges-in-representation-learning-multi-modal-learning`

[8] `http://www-etud.iro.umontreal.ca/~goodfeli/mlc2013.html`

of the images and using their labels as the incorrect label for the remaining labels. This removes the matching aspect of the problem and forces the classifier to label each image independently.

5 Advice to Contest Organizers

Organizing a contest requires a significant amount of work from all parties involved. We offer some suggestions for running a succesful contest:

Allocation of Time: Budget time for the following tasks: *Before the contest launches:* Creation of new datasets, verification that state of the art algorithms perform well but have room for improvement on the dataset, preparation of baseline solutions, design of rules for the contest. *During the contest:* Fielding questions (on contest rules, how to use the contest website, etc.), resolving portability issues with contest baselines. *After the contest* Verification of the winners' submissions, distributing the private test data, preparing presentations and papers about the contest.

Designing Rules: Some things to consider: Should "transductive" methods that are allowed to observe all test set inputs be allowed? Are contestants prohibited from labeling the public leaderboard test data and training or cross-validating with it? What about training with outside data, or scraping the web for higher resolution versions of input images? How will you enforce the rules? Datasets that humans can label present many difficulties. Remember that you need to prevent not just training on the test set, but also selecting hyperparameters on it. The best way to do this is to require all entrants to upload their trained models at the end of the contest. The organizers release the test set only after all models are frozen. Entrants then run their submission on the test set and upload the predictions . The organizers then verify that the winning submissions' predictions were indeed generated by the previously uploaded model. Using this system is a powerful deterrent to cheating. In order to run the contest smoothly, it is important to plan these measures in advance and put them in the rules from the start. We initially had fewer cheating deterrants in place, expecting only a small number of competitors from the academic deep learning community, but within days of launching the contest someone had already hand-labeled the entire public test set for the multimodal learning contest. Note that contestants are interested in obtaining a high rank on the leaderboard even if they do not win a prize (on Kaggle, one can earn "Kaggle points" for placing in the top 10% or 25% of a contest). It's important to reserve the right to verify all submissions and remove leaderboard entries that can't be verified.

Difficulty and Participation Rate: Err on the side of making the contest too hard rather than too easy. We erred on the side of making the contests easy, in order to increase participation, and this made the multimodal contest too easy to be interesting. While past workshop-based contests have had a low participation rate (example: 4 teams in the NIPS 2011 transfer learning challenge [22, 23]) the participation rate problem can be completely solved by hosting the contest on Kaggle. Even our least popular challenge had 26 teams.

Organize Multiple Contests Simultaneously. The marginal cost of running a second or third contest is low compared to the fixed cost of launching one contest, and the additional contests provide some insurance that you will obtain interesting results even if one contest turns out to be poorly devised.

Provide Baselines and a Leaderboard. Baselines boost participation since entrants don't need to write boilerplate code to load the data, etc.

6 Discussion and Conclusion

Competitions offer a different and important viewpoint on machine learning algorithms than research papers do. Research papers are expected to be extremely novel. When writing research papers, the most talented machine learning practitioners focus their skills on tuning methods that they themselves invented. Contests offer the opportunity to see what happens when a different incentive structure is applied: skilled practitioners use whatever means they think will help them win, regardless of how novel the method is or whether they invented it. The use of a completely new test set also makes the results of the contest a more realistic evaluation of generalization error. When interpreting the results of a contest, it is important to remember that a contest is not a controlled experiment complete with statistical analysis. However, contests can serve to refocus our attention on algorithms that perform well, but may not otherwise receive their due attention in the research community. This year's contest highlighted the performance of SVM loss functions, sparse filtering, and entropy regularization. We hope these results help machine learning practitioners improve the performance of their algorithms, and that future contest organizers are able to use this report to plan more contests that highlight more effective algorithms.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. Technical Report arXiv:1206.5538, U. Montreal (2012), `http://arxiv.org/abs/1206.5538`

[2] Goodfellow, I.J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., Bengio, Y.: Pylearn2: A machine learning research library. arXiv preprint arXiv:1308.421 (2013a)

[3] Guyon, I., Dror, G., Lemaire, V., Taylor, G., Aha, D.W.: Unsupervised and transfer learning challenge. In: Proc. Int. Joint Conf. on Neural Networks (2011)

[4] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: Deep Learning and Unsupervised Feature Learning Workshop, NIPS (2011)

[5] Ngiam, J., Koh, P.W.W., Chen, Z., Bhaskar, S.A., Ng, A.Y.: Sparse filtering. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P., Pereira, F.C.N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 24, pp. 1125–1133 (2011)

[6] Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

[7] Cortes, C., Vapnik, V.: Support vector networks. Machine Learning 20, 273–297 (1995)

[8] Romaszko, L.: A deep learning approach with an ensemble-based neural network classifier for black box icml 2013 contest. In: Workshop on Challenges in Representation Learning, ICML (2013)

[9] Lee, D.-H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML (2013)

[10] Grandvalet, Y., Bengio, Y.: Semi-supervised Learning by Entropy Minimization. In: NIPS 2004. MIT Press, Cambridge (2005)

[11] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and composing robust features with denoising autoencoders. In: ICML 2008 (2008)

[12] Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. In: ICML (2013b)

[13] Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)

[14] Susskind, J., Anderson, A., Hinton, G.E.: The Toronto face dataset. Technical Report UTML TR 2010-001, U. Toronto (2010)

[15] Bergstra, J., Cox, D.D.: Hyperparameter optimization and boosting for classifying facial expressions: How good can a "null" model be? In: Workshop on Challenges in Representation Learning, ICML (2013)

[16] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36(4), 193–202 (1980)

[17] Lowe, D.: Object recognition from local scale invariant features. In: ICCV 1999 (1999)

[18] Tang, Y.: Deep learning using linear support vector machines. In: Workshop on Challenges in Representation Learning, ICML (2013)

[19] Ionescu, R.T., Popescu, M., Grozea, C.: Local learning to improve bag of visual words model for facial expression recognition. In: Workshop on Challenges in Representation Learning, ICML (2013)

[20] von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2004, pp. 319–326. ACM, New York (2004)

[21] Feng, F., Li, R., Wang, X.: Constructing hierarchical image-tags bimodal representations for word tags alternative choice. In: Workshop on Challenges in Representation Learning, ICML (2013)

[22] Le, Q.V., Ranzato, M., Salakhutdinov, R., Ng, A., Tenenbaum, J.: NIPS Workshop on Challenges in Learning Hierarchical Models: Transfer Learning and Optimization (2011), `https://sites.google.com/site/nips2011workshop`

[23] Goodfellow, I., Courville, A., Bengio, Y.: Large-scale feature learning with spike-and-slab sparse coding. In: ICML (2012)

A Comparative Study on Image Retrieval Systems

Chee Sheen Chan and Jer Lang Hong

School of Computing and IT, Taylor's University
{cheesheen.chan,jerlang.hong}@taylors.edu.my

Abstract. Search Engines such as Google is capable of processing user queries in a fast and efficient way. Though they are proven to be reliable, they are still incapable of handling complex queries. For example, Image Search Engines generally suffer from low accuracy when handling complex queries due to the lack of information available in an image. This paper presents the various image retrieval systems available currently, with in depth discussions on their operation and drawbacks. We also demonstrated the potential of using ontological technique in image retrieval systems, which has shown promising results in many research domains.

Keywords: Human-Centred Design, Webpage Segmentation, Image Indexing, Search Engines.

1 Introduction

With the wide availability of high speed networks and the growth of World Wide Web, information has become widely accessible in the world. Users can easily acquire any information in any location of their choice using devices such as smart phones and PDAs. There is a wide range of information available in the web; the most typical ones are text based information. Recently, images have become an indispensable method for humans to express their thoughts and needs. With the recent social networking sites and microblogging, people upload and share their images or other multimedia-related content over the world on a global scale. Images can be considered as one of the vital medium of expressing thoughts.

Current commercial web-based image search engines such as Google Images, Bing, Yahoo, etc, provide image annotation so that users can search for them using search queries. However, these search engines are unable to handle complicated search queries especially search queries which are too specific. Such systems work generally fine for most simple queries (e.g., apple, cat, watch, etc). Complicated queries such as "Amazon Rainforest on Fire" are also common and are the reason behind most of these failed searches.

The accuracy of image search engines is crucial and highly dependent on the image annotation. Fortunately, images in web do come with precious contextual information. This information is widely available and can be located within the nearby region of that particular image. However, not all the information located nearby are relevant to the image, likewise not all the information located far away are irrelevant to the im-

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 125–132, 2013.

age. The focus of this study is thus to garner an in-depth understanding of this information for indexing purposes.

To provide an in depth understanding of this study, our paper is written into two parts. The first part of this paper describes the type of image retrieval systems available currently, with their pros and cons provided. The second part of this paper provides an overview of the different categories of image retrieval systems. While the study of image retrieval systems is a huge and wide area, our study in this paper is primarily focused on the state of the art image retrieval systems and also on the future trends of developing these systems.

This paper contains several sections. Section 2 describes the various image retrieval systems while Section 3 gives the categorization of the image retrieval systems. And last but not least, Section 4 summarizes our work.

2 Image Retrieval Systems

An image retrieval system can be defined as a computer system for browsing, searching and retrieving images from a large database of digital images. Most traditional and common approaches of image retrieval utilize variousmethods of adding metadata such as captions, keywords, or descriptions to the images so that retrieval can be performed over the annotation words.

There are two aspects to an image retrieval system; indexing and searching. The metadata of the image is indexed and stored in a large database and when a search query is performed, the image search engine looks up the index, and queries are matched with the stored information. The results are presented in ascending order according to its relevancy.

Generally, the image retrieval system can be divided into 2 methods, which are content-based and concept based. Relatively, a third method has become available as a result of recent development and research.

Content-Based Image Retrieval Systems (CBIR)

Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision techniques to the image retrieval problem, which is, the problem of searching for digital images in large databases. It is not required for images to be indexed with textual labels for these systems. Instead, images are indexed and retrieved according to their low-level visual features of colors, shapes, texture, and/or spatial information spatial information [4], [6].

There is a growing interest in Content-based IR due to the limitations inherent in metadata-based systems, as well as a wide spectrum of potential uses for efficient image retrieval. Textual information about images can be easily searched using existing technology; however human intervention is required to manually describe every single image in the database. This is impractical for very large databases, or for images that are generated automatically, e.g. from surveillance cameras. It is also possible to miss images that use different synonyms in their descriptions. A disadvantage of Content-based IR is the method of performing the query, since images are indexed with low-level features, users can only perform queries by sketch, color composition and/or example (i.e. by providing an example image). Query by sketch method is not desirable for non-artistically inclined users and query by example method assumes that the user already has an example on hand. These drawbacks have limited CBIR to domain-specific applications and words remain predominant in image indexing and retrieval systems, at least in the foreseeable future [17].

Concept-Based Image Retrieval Systems

Concept-based image indexing and retrieval, also variably named as "description-based" or "text-based" image indexing/retrieval, refers to retrieval from text-based indexing of images that may employ keywords, subject headings, captions, or natural language text. It is opposed to Content-based image retrieval. “Cats”, “oranges”, “accident on the highway” and etc are examples of such indices. Legacy keyword-based image retrieval systems, CBIR systems with high-level semantics, as well as Web-based image retrieval systems could be categorized under concept-based image retrieval systems.

Legacy keyword-based IR systems involve tagging the images manually with a keyword, which initially became costly and impractical. However, with Web 2.0 technology, it became possible to easily tag images manually with a keyword without the drawbacks as implemented in many sites such as Flickr, Deviantart, and similar image sharing sites. However, despite such tagging methods, this approachis proven have its drawbacks in that it is highly dependent on the indexer; the annotations are error-prone, incomprehensive and that its range of successful queries is onlyrestricted to the interpretation of the indexer/annotator [4], [8], [20].

Concept-based IR systems with higher semantics are capableof learning high-level semantic concepts from low-level visual features using advance computer vision and machine learning techniques, concentrating on reducing the semantic gap; such systems include semantic-based image retrieval systems [1], [4], [13] and signal/semantic-based system [15]. Bradshaw [1] attempted to use a probabilistic model in his semantic-based system to recognize four concepts (*i.e.* natural/man-made and indoor/outdoor) in an image. Jeon et al. [9] on the other hand implemented a cross-media relevance model which could identify 70 object concepts, while Li et al. [13] cross-media relevance model could identify 101 object concepts. However, these implementations comes with a drawback, visual information is lost in the learning process. This is where the signal/semantic-based systems come in, to address this issue. In Liu et al. survey on Concept-based IR with high-level semantics, it is mentioned that Li et al. [13] 101 object concepts form the largest vocabulary set used in object recognition. However, this is far from the 30,000 object concepts perceived by humans which show that there is still much to do in order to fill the semantic gap.

Multi-modal Web-Based Image Retrieval Systems

Both Context-based and Concept-based systems have their pros and cons. And as such, many researchers have delved into the idea of exploiting the strengths of these two systems through fusion and multi-modality, all the while addressing the weaknesses that both systems had.

There are many research papers concerning the fusion of the visual features of the Concept-based systems and the contextual information of images on the internet [14], [22]. These researches focus on drawing upon the contextual information as a source of countless semantic concepts to increase the cardinality of the sets of semantic classes, and attempting to overcome the poor retrieval performance of current web-based systems with the image visual content.

Some research combined both textual and visual information to cluster images rather than annotate them [22]. Other research focused on the fusion of both sources for image annotation purposes [14], [18], [22]. In the earlier stages of this research [18], [14], the textual and visual information were kept in separate data repositories and loosely coupled by means of relevant feedback from the user. Users were required to provide feedback on which images were relevant/irrelevant in the search result returned by their keyword-based query; though this method causes the system to appeal less attractively to the users. As researchers continued refining the methods, some research strongly coupled both textual and visual image content by using a bootstrapping approach and graph learning method. The system recall rate has been improved, which effectively eradicates the need for user feedback; however, only a minor increase is observed in the system retrieval precision. A probable explanation for this is the poor semantic relevance of the textual information attached to the image, which only raises the concern of the original problem; the unprecedented quality of the contextual information of a web image faced by current Web-based systems.

3 Categorization of Image Retrieval Systems

These systems can be categorized into two systems; pure text-based systems or multi-modal systems (aka fusion-based models). In this section, an analysis shall be made on how these two systems make use of an image's contextual information.

Text-Based Image Retrieval Systems

Most commercial image search engines – Google Image, Yahoo! Image and Bing Image (formerly Live Search Image and MSN Image) , are text-based systems widely used by the public to search for images on the internet. These three image search providers originated from general text information search engines (Google, Yahoo, Altavista, etc).

Such systems typically rely on text to index images on the internet. These keyword-based systems make use of the image's filename, the hyperlink text pointing to the image, and/or the text adjacent to or surrounding the image. In 2009, Google included an image content-based similarity feature to look up other similarly colored or textured images after the initial textual query. This feature shares some similarities to the early fusion-based models that loosely couple textual and visual features where users can submit a text query, select and view other visually similar images from the query result [14], [18], [21].

While it is hard to determine how much text these search engines consider as adjacent to/surrounds an image, it has been reported by Feng et al. [6] that the first or last 32 words in the text nearest to an image appears to be most descriptive of the image according to a survey conducted by Google. However, in the case of Bing Image Search (Microsoft), a section of text is extracted using Microsoft's patented webpage segmentation algorithm, otherwise known as VIPS which partitions a webpage into several smaller semantic blocks rather than considering a number of terms as the text surrounding the image (like Google's method), instead.

Examples of pure text-based systems can be found in [4], [7], [8], [11], [19]. An image representation model called *Weight ChainNet* is introduced [29]. This model is based on a lexical chain. The Image filename, image ALT, page title and image caption (e.g. the entire paragraph containing the image) are considered as part of the image's contextual information, and these texts are modeled as different lexical chains in a Weight ChainNet model. The best performance has been demonstrated by a proper combination of these chains, each with their own appropriate weightages.Several tests of different weight combination are performed to obtain the optimized weight for each chain. Hence, the query results is further refined through relevance feedback methods.

Image contextual information was considered as text from multiple sources as well, with each part of the text being regarded as an independent source of evidential information. Four possible sources of evidence are proposed: description tags, meta tags, full text and text passages (i.e. surrounding text – words located close to the images). An initial experiment was conducted to determine the best size of text passages where 5, 10 and 20 terms before and after an image as well as full text were tested. It can be observed that the size of 20 terms gives the best result. These sources of text are then combined in a Bayesian network model to improve the retrieval quality of image retrieval systems. A combination of text passages and description tags provides the best retrieval results and poor retrieval is shown when these text sources are used in isolation.

An image's contextual information can range from the image caption (e.g. a paragraph of text) to the entire article text body [8]. HTML tags are classified into either *block* or *style* tags. The page layout or the relative positioning of the content is affected by the block tag element whereas the visual attribute of the content such as font size or color is affected by the style tag element. The identified block tag is then rendered into a content block on the webpage, which they consider as the article text body (i.e. the main content of the webpage). A linguistic-based semantic similarity algorithm is applied to associate the image to the article text body. A match is found for the named entities between the image caption and the text from each of the content blocks that make up the article text body. Though, images without captions are considered as non-article images.

The last two papers concentrate on indexing images solely on the surrounding text extracted from the entire webpage content. According to Gong et al. [7], the image context is put through a stop word removal and stemming process, and the result is partitioned into 3 groups – page-oriented text, TM (texts from the title and meta tags), link-oriented text, LT (texts attached to the image tag) and caption-oriented text, BT (texts of the body). For each term in a block, a local weight corresponding to its semantic relevance to the image is calculated based on its local occurrence (using *tf-idf*weighting model) and distance of the block to the image. Thus, the overall relevance of a term to an image is determined as the sum of all its local weightages multiplied by the corresponding distance factors, in an attempt to rank relevant terms higher than irrelevant terms.

Multi-modal Systems

In the earlier years where loosely-coupled models utilized both image contextual information and image content, WebSeer [5], AMORE [16], ImageRover[18] and iFind[2] all focused on using textual cues that come from the image's filename, ALT attribute within the <IMG> tag of the HTML file, link text, title of the HTML page and the text surrounding the image (the definition for this differs for every system). According to Sclaroff et al. [18], the text surrounding the image is defined as 10 words appearing before the <IMG> tag and 20 words appearing after the <IMG> tag. Emphasized words in bold and italics, word frequency and word proximity to the image are all taken into account using the Latent Semantic Indexing (LSI) method. The words that are closer to the image are ranked higher for word proximity. Some resear consider a paragraph of text withoutspecifying a number of words before and after an image [5], [16]. Frankel et al. [5] consider texts are weighed according to the text source/location of the image filename, image caption, ALT attribute, HTML page title and link text; where these features are assumed to be analogous to the likelihood of the text being useful in an image search, though Chen et al. [2] use*tf-idf* model torank the texts. On the other hand, the hyperlink text and html addresses (e.g. uniform resource locator – URL) is being utilized by inWebSeek image and video retrieval system to index multimedia resources on the internet [21].

4 Conclusions

The different types of image retrieval systems, their differences and drawbacks have been addressed in this paper. Content based image retrieval systems require human annotation and labeling while context based image retrieval systems require surrounding contextual information. There are problems with both, the former requires extensive labor while the latter can be automated by a well-designed computer system. Based on our observations, it seems that current trends concentrate more on the multi modal approach, and with the introduction of ontologies, current approaches tend to adopt these techniques as a tool for image indexing and retrieval. Though ontology techniques are generally slow, it is foreseen that these techniques will be the future tools due to their ability to analyze the conceptual and semantic properties of a document.

References

1. Bradshaw, B.: Semantic Based Image Retrieval: A Probabilistic Approach. In: Proceedings of the 8th Annual ACM International Conference on Multimedia, pp. 167–176. ACM (2000)
2. Chen, Y., Ma, W.Y., Zhang, H.J.: Detecting web page structure for adaptive viewing on small form factor devices. In: Proceedings of the 12th International Conference on World Wide Web, pp. 20–24 (2001)
3. Coelho, T., Calado, P., Souza, L., Ribeiro-Neto, B., Muntz, R.: Image retrieval using multiple evidence ranking. IEEE Transactions on Knowledge and Data Engineering 16(4), 408–417 (2004)
4. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Dom, B., Gorkani, M., Hafner, J., et al.: Query by image and video content: the QBIC system. Computer 28(9), 23–32 (1995), doi:10.1109/2.410146
5. Frankel, C., Swain, M., Athitsos, V.: Webseer: an image search engine for the world-wide web. Tech Rep. 94–14 (1996)
6. Gevers, T., Smeulders, W.: Color constant ratio gradients for image segmentation and similarity of texture objects. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. 1–18. IEEE (2001)
7. Gong, Z., Hou U, L., Cheang, C.W.: Web image indexing by using associated texts. Knowledge and Information Systems 10(2), 243–264 (2006)
8. Inoue, M.: On the need for annotation-based image retrieval. In: Proceedings of the Workshop on Information Retrieval in Context (IRiX), Sheffield, UK, pp. 44–46 (2004)
9. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 119–126 (2003)
10. Joshi, P.M., Liu, S.: Web document text and images extraction using DOM analysis and natural language processing. In: Proceedings of the 9th ACM Symposium on Document Engineering - DocEng 2009, p. 218. ACM Press, New York (2009)
11. Leong, C.W., Mihalcea, R., Hassan, S.: Text Mining For Automatic Image Tagging. In: COLING, pp. 647–655 (2010)
12. Leong, C.W., Mihalcea, R.: Explorations in Automatic Image Annotation using Textual Features. In: Linguistic Annotation Workshop, pp. 56–59 (2009)
13. Li, F.-F., Fergus, R., Perona, P.: Learning generative visual models from few train-ing examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106(1), 59–70 (2007)
14. Lu, Y., Hu, C., Zhu, X., Zhang, H.J., Yang, Q.: A unified framework for semantics and feature based relevance feedback in image retrieval systems. In: Proceedings of the 8th Annual ACM International Conference on Multimedia, pp. 31–37. ACM (2000)
15. Meghini, C., Sebastiani, F., Straccia, U.: A model for multimedia information retrieval. Journal of the ACM (JACM) 48(5) (2001)
16. Mukherjea, S., Hirata, K.: Amore: A World Wide Web image retrieval engine. World Wide Web 2, 115–132 (1999)
17. Rorissa, A.: User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. Information Processing & Management 44(5), 1741–1753 (2008)

18. Sclaroff, S., Taycher, L., Cascia, M.L.: ImageRover: A content-based image browser for the World Wide Web. In: IEEE Workshop on Content-Based Access of Image and Video Libraries, pp. 2–9 (1999)
19. Shen, H.T., Ooi, B.C., Tan, K.-L.: Giving meanings to WWW images. In: Proceedings of the 8th Annual ACM International Conference on Multimedia, pp. 39–47 (2000)
20. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
21. Smith, J.R., Chang, S.F.: An image and video search engine for the world-wide web. In: Symposium on Electronic Imaging Science and Technology-Storage & Retrieval for Image and Video Databases V (1997)
22. Wang, C., Zhang, L., Zhang, H.-J.: Learning to reduce the semantic gap in web im-age retrieval and annotation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 2008, p. 355. ACM Press, New York (2008)

Identifying Association between Longer Itemsets and Software Defects

Zeeshan A. Rana, Sehrish Abdul Malik, Shafay Shamail, and Mian M. Awais

LUMS School of Science and Engineering Lahore, Pakistan
{zeeshanr,11030025,sshamail,awais}@lums.edu.pk

Abstract. Software defects are an indicator of software quality. Software with lesser number of defective modules are desired. Prediction of software defects using software measurements facilitates early identification of defect-prone modules. Association relationship between software measures and defects improves prediction of defective modules. To find association relationship between software measures and defects, each numeric measure is divided into bins. Each bin is called 1-itemset (or an itemset of length 1). When certain itemsets and defective modules appear together in a dataset, they are considered associated with each other. Frequency of their co-occurrence depicts the strength of the association relationship. Existing studies find the relationship between 1-itemsets and defective modules. Itemsets that have high association with defects are called focused itemsets. Focused itemsets can be used to build prediction models with higher *Recall* values. This paper explores the relationship between defective modules and itemsets with length greater than 1. Focused itemsets with length greater than 1 involve multiple bins at same time. Identification of the focused itemsets has improved the performance of decision tree based defect prediction model.

1 Introduction

Software quality is an important characteristic for success of a software system. One way to measure software quality is to count number of defective modules. Prediction of defective modules helps in: a) planning of resources during development and testing b) reducing defect correction cost and time [8], [9], [10], [12]. Techniques used to predict defects include statistical methods, machine learning and data mining methods, parametric models and mixed algorithms [4]. Most of the defect prediction techniques use software measures collected during design and coding phases. Studies have emphasized the need to understand relationship between software measures and defects [5], [12]. Association mining have been used to find association relationship between the two [12]. The association relationship is identified by discretizing the numeric software measures into bins and identifying 1-itemsets that coexist with software defects [12]. 1-itemset means one interval from the range of values of a single software measure (also known as an attribute). Each attribute is divided into multiple bins and software defect data contains multiple attributes. Therefore association information between one bin of a single attribute and defective modules is not enough. Also, to

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 133–140, 2013.

understand the relationship in a holistic manner [5] requires the contribution of multiple bins in the relationship. To get a more complete view of the relationship this paper identifies the longer itemsets that co-occur with defective modules. The longer itemsets provide more information about the defective modules than the information provided by 1-itemset. This paper also uses these itemsets to develop decision tree based prediction model [6]. The itemsets have improved the *Recall* (or true positive rate) of the decision tree model.

The rest of this paper is organized as follows. Section 2 presents an overview of the related work. Section 3 presents our research methodology. Results are presented in section 4 and section 5 discusses these results. Finally, section 6 concludes the paper.

2 Related Work

Among other techniques data mining and machine learning methods are widely used for defect prediction. Challagulla et. al have compared the linear regression, pace regression, support vector regression, logistic regression, neural network, naive bayes, instance-based learning, J48 trees and 1-rule using four defect data sets [4]. Their findings are that instance-based learning together with 1-rule gave better prediction and that size and complexity attributes are not enough for accurate prediction. Defect prediction models have been criticized for using size and complexity measures only by Fenton et. al [5]. Fenton et al advocate use of a model with a holistic view of a software system to predict defects. Unlike Challagulla, Fenton et. al do not use static code measures and suggest a model based on Bayesian Belief Networks (BNN) which involves expert based judgement. However, there are numerous studies that use static code attributes to predict defects. Menzies et. al stated that predictor's performance could have suffered if static code attributes were not helpful [8]. Their motivation was to utilize code metrics to improve the performance of defect predictors. Based on code attributes Ma et al. suggested using association based classification for defect prediction [2]. They have compared association based classification, CBA2, with other rule based classification methods. Their comparison suggested that CBA2 performed better than C4.5 and RIPPER based on the measures Area Under Curve(AUC), accuracy, sensitivity and specificity. Kamei et. al have proposed an association mining and logistic regression based approach to predict defect prone modules [7]. The hybrid approach was superior to other prediction models on the bases of $lift$ measure [6], [7]. Other studies have also used the static code attributes and association mining to facilitate prediction of software defects [1], [12]. Zafar et. al have identified attribute values that have high association with defective modules [12].

3 Methodology

This paper adapts the methodology by Zafar et. al [12] to explore the relationship of itemsets with length ≥ 2 with defective modules. The public data [3]

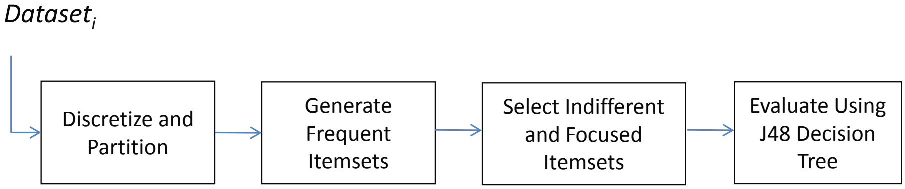

Dataset$_i$
Discretize and Partition
Generate Frequent Itemsets
Select Indifferent and Focused Itemsets
Evaluate Using J48 Decision Tree

Fig. 1. Methodology to Find Focused Itemsets

used for this study consists of numeric attributes. These numeric attributes are software measures collected during development phases. The data also has a class attribute. Each record in a dataset represents a software module. The class attribute categorizes each module as defective D or not-defective ND. Figure 1 shows the steps involved in finding the longer focused itemsets. As shown in the figure the first step is to discretize date so that itemsets can be generated in later phases of the methodology. Datasets used are imbalanced and consist of more ND modules than the D modules [12]. Each dataset is partitioned into two based on the class attribute value. For each partition Apriori algorithm [6] is used to generate frequent itemsets of length two and greater. Support of all the itemsets is used to identify focused and indifferent itemsets. J48 [11] (java implementation of C4.5 [6]) is used to validate the generation focused itemsets. Rest of the section discusses each step in detail.

3.1 Discretization and Partitioning of Data

Apriori algorithm works on discrete data only so each attribute is divided in 10 equi-frequency bins. From these bins, combinations of bins that are highly associated with defective modules are identified. The discretized data is partitioned such that partition D_t includes defective modules whereas, partition D_f includes not-defective modules. As a next step, Apriori algorithm is applied to generate frequent itemsets.

3.2 Finding Frequent Itemsets and Support of Each Itemsets

Frequent itemsets are generated for each partition with respect to their association with the class attribute. The frequent itemsets also satisfy a minimum support threshold in addition to occurring frequently with the class attribute in a partition. Apriori algorithm generates itemsets of various lengths in each partition. The 2-itemsets and 3-itemsets studied in this paper satisfy the minimum support thresholds $MinSupport_t$ and $MinSupport_f$ in partitions D_t and D_f respectively. Support of an itemset is the proportion of the modules in a partition that contain the itemset.

3.3 Selection of Indifferent and Focused Itemsets

Indifferent itemsets are the itemsets that appear in both partitions D_t and D_f and satisfy α_t and α_f thresholds in the respective partition. These itemsets do not affect the classification of D modules and can be ignored when developing a classification model with high *Recall*. Itemsets that appear in the partition D_t, satisfy α_t and are not indifferent itemsets are called focused itemsets. Focused itemsets are the combinations of bins that co-occur with the defective modules.

3.4 Evaluation Framework

This paper uses decision trees based 2 phase criterion to evaluate results. In phase 1, J48 decision tree is generated for each datasets and the branches of the tree that trace towards defect prone modules are analyzed. From these branches the attributes that appear at decision points are observed. These attributes with their ranges are compared to the focused 2 and 3-itemsets and results are validated. If a certain combination of itemsets appears as decision nodes in more datasets, vote count for the combination will be high. High vote counts represent that certain combinations of bins associate highly with defects and should be considered important when developing prediction models with better *Recall*. *Recall* (or True Positive Rate) is proportion of actual D modules in the modules predicted as D.

In phase 2, the impact of indifferent and focused itemsets on detection of defects is studied. Performance of the decision tree is measured at 4 different points: without any pre-processing, after dropping the attributes with indifferent itemsets, after relabeling focused 1-itemsets, after relabeling focused 2-itemsets. The focused 1-itemsets and 2-itemsets are relabeled as missing value for the ND modules only. This process of relabeling the focused itemsets should improve *Recall* of the decision tree. Increase in *Recall* should indicate that these itemsets are related to the defective modules.

4 Results

To find focused itemsets, each dataset passes through the phases described in section 3. Four datasets [3] used in this study are listed in Table 1. While passing

Table 1. Min Support, α_t and α_f, used in this study, for each dataset

	Partition D_t		Partition D_f	
Dataset	$MinSupport_t$	α_t	$MinSupport_f$	α_f
cm1	15 %	25 %	10 %	30 %
jm1	15 %	20 %	10 %	30 %
kc1	15 %	25 %	20 %	25 %
kc2	20 %	30 %	20 %	30 %
pc1	20 %	25 %	10 %	30 %

through the phases all independent numeric attributes of each dataset are discretized into 10 equi frequency bins. The discretized dataset is partitioned into D_f and D_t. Apriori algorithm is applied on each partition of all the datasets to generate frequent itemsets. Minimum support thresholds $MinSupport_t$ and $MinSupport_f$ used to apply Apriori are given in Table 1. As the next step support and frequency of all the 2-itemsets and 3-itemsets in all the datasets is calculated. Indifferent and focused itemsets are identified using the α_t and α_f threshold, given in Table 1. All the itemsets are sorted in descending order according to their support value. The itemsets above the support threshold value are marked either as focused or indifferent. As in the study by Zafar et al [12] $\alpha_t \leq \alpha_f$ for all datasets. This is because the data is imbalanced and partition D_t contains very small number of examples as compared to examples in partition D_f.

Table 2 shows top 3 focused and indifferent 2-itemsets with their support. Itemsets in bold face are focused itemsets and indifferent itemsets have been marked as {Indifferent}.

Table 2. Top 3 2-Itemsets and their $Support_i$ in each partition

	Partition D_t		Partition D_f	
Dataset	2-Itemset	$Support_i$	2-Itemset	$Support_i$
$CM1$	ev(g)= '(-inf-1.2]' locCodeAndComment='(-inf-0.5]' {Indifferent}	61.22 %	ev(g)= '(-inf-1.2]' locCodeAndComment='(-inf-0.5]' {Indifferent}	76.61 %
	loc=‘(65.5-inf)’ locCodeAndComment =(-inf-0.5]	34.69 %	iv(g)= '(-inf-1.2]' locCodeAndComment='(-inf-0.5]'	50.78 %
	loc=‘ (65.5-inf)’ lOComment=‘(34.5-inf) ’	28.57 %	ev(g)= '(-inf-1.2]' iv(g)= '(-inf-1.2]'	46.77 %
$JM1$	lOComment='(-inf-0.5]' locCodeAndComment='(-inf-0.5]' {Indifferent}	49.72 %	lOComment='(-inf-0.5]' locCodeAndComment='(-inf-0.5]' {Indifferent}	66.76 %
	lOBlank='(-inf-0.5]' locCodeAndComment ='(-inf-0.5]'	21.56 %	iv(g)='(-inf-1.2]' locCodeAndComment='(-inf-0.5]'	37.65 %
	e='(48232.24-inf)' t='(2679.57-inf)'	21.27 %	ev(g)='(-inf-1.2]' iv(g)='(-inf-1.2]'	35.39 %
$KC1$	ev(g)='(-inf-1.2]' locCodeAndComment='(-inf-0.5]' {Indifferent}	65.34 %	ev(g)='(-inf-1.2]' locCodeAndComment='(-inf-0.5]' {Indifferent}	86.26 %
	e='(14140.38-inf)' t='(775.605-inf)'	28.83 %	ev(g)='(-inf-1.2]' iv(g)='(-inf-1.2]'	66.74 %
	n='(147.5-inf)' v='(795.61-inf)'	28.53 %	iv(g)='(-inf-1.2]' locCodeAndComment='(-inf-0.5]'	65.68 %

Continued on next page

Table 2 – continued from previous page

	Partition D_t		Partition D_f	
Dataset	2-Itemset	$Support_i$	2-Itemset	$Support_i$
$KC2$	ev(g)=’(-inf-1.2]’ lOCodeAndComment=’(-inf-0.5]’ {Indifferent}	44.86 %	ev(g)=’(-inf-1.2]’ lOCodeAndComment=’(-inf-0.5]’ {Indifferent}	83.13 %
	v=’(1403.34-inf)’ uniq_Opnd=’(36-inf)’	34.58 %	ev(g)=’(-inf-1.2]’ lOComment=’(-inf-0.5]’	70.84 %
	uniq_Opnd=’(36-inf)’ total_Op=’(151.5-inf)’	34.58 %	lOComment=’(-inf-0.5]’ lOCodeAndComment=’(-inf-0.5]’	69.40 %

5 Analysis and Discussion

Most of the itemsets found in [12] appear in longer itemsets generated in this study. Almost all the datasets contain the 1-itemset of study [12]in combination with other attributes as focused 2 and 3-itemsets with the exception of jm1 dataset. For example, for cm1 dataset loc= ’(65.5-inf)’, lOComment= ’(34.5-inf)’, and n= ’(400.5-inf)’ appeared as focused 1-itemset [12]. In this study, 57% of the focused 2-itemsets and 80% of the focused 3-itemsets contain these 1-itemsets in combination with other attributes for cm1 dataset. This shows that 1-itemsets that appeared as focused earlier still appears as focused but now in conjuncture with other attributes.

Also, it can be analyzed that 2-itemsets and 3-itemsets frequently contain the same attributes with same ranges but in different combination. For example the focused 2-itemsets for pc1 dataset are: loc=’(54.5-inf)’ N=’(279.5-inf)’, loc=’(54.5-inf)’ V=’(1713.03-inf)’, loc=’(54.5-inf)’ B=’(0.565-inf)’, loc=’(54.5-inf)’ lOCode=’(54.5-inf)’. 3-itemsets for the same dataset are: loc=’(54.5-inf)’ N=’(279.5-inf)’ V=’(1713.03-inf)’, loc=’(54.5-inf)’ N=’(279.5-inf)’ B=’(0.565-inf)’, loc=’(54.5-inf)’ N=’(279.5-inf)’ uniq_Opnd=’(50.5-inf)’. It is clear that the combination of loc=’(54.5-inf)’, N=’(279.5-inf)’, V=’(1713.03-inf)’ and B=’(0.565-inf)’ primarily contribute towards defects. These measures with the ranges and combinations as identified by the experiment can be used by the managers to monitor the quality of the software project.

The identified indifferent itemsets show lower values of attributes for all the datasets meaning thereby low values of the 2-itemset do not contribute in occurrence or absence of defects. Most of the 2-itemsets in all datasets are similar. However, lOBlank=’(-inf-0.5]’ locCodeAndComment=’(-inf-0.5]’ appeared in $jm1$ only. This shows that low number of blank lines and small sized code do not affect the module as being defective or non-defective. Similarly, ev(g)=’(-inf-1.2]’ lOComment=’(-inf-0.5]’ locCodeAndComment=’(-inf-0.5]’ is the 3-itemset that is common in most of the datasets as indifferent. The number of indifferent 3-itemsets reduces in all the datasets because the combination of attributes becomes rare in the partitions D_t and D_f.

The results of the experiment are validated by generating decision tree for each dataset using Weka. For CM1 decision tree the root node lOComment=’(34.5-inf)’

Table 3. Performance of Decision Tree Model in terms of *Recall*

	$KC1$	$KC2$	$JM1$	$CM1$
No Pre Processing	0.34	0.505	0.13	0.22
Attributes Dropped	0.507	0.527	0.155	0.154
1-Itemsets Relabelled	0.553	0.587	0.258	0.449
2-Itemsets Relabelled	0.553	0.613	0.339	0.449

together loc='(65.5-inf)', ev(g)='(-inf-1.2]', and lOBlank = '(-inf-0.5]' contributes towards defective modules. Our results are similar to these results that is combination of these attributes exist in 2-itemsets and 3-itemsets.

Performance of the Decision Tree (DT) model (in terms of *Recall*) has increased by identification of the focused and indifferent itemsets. Table 3 shows *Recall* of the decision tree at four points. First row shows the performance of the DT when no modifications are done in data. Row 2 shows the performance when attributes with indifferent itemsets are dropped from the data. Row 3 and 4 respectively show the impact of relabeling 1-itemsets and 2-itemsets as missing values. *Recall* of the model has increased for each dataset after each step. Increase in performance has been 61. 76%, 21.38%, 160.76% and 100.04% for the datasets $KC1, KC2, JM1$, and $CM1$ respectively.

6 Conclusions and Future Work

Information regarding association between software measurements and defective modules is useful in predicting defective modules early in lifecycle. This paper identifies the association relationship between defective modules and combination of 2 and 3 software measures. The paper divides the numeric software measures into equi frequency bins (called itemsets) and uses Apriori algorithm to see which longer itemsets (with length ≥ 2) associate with defective modules. These longer itemsets are called focused itemsets. Most of the longer itemsets include the 1-itemsets reported in [12]. J48 decision trees have been developed for each of the discretized datasets. The branches of the tree tracing towards defective modules have been examined to validate if the longer itemsets appear as decision nodes in the tree. For all the datasets the focused itemsets appear as decision nodes of the respective trees. Focused itemsets have been used to develop decision tree based prediction model. Average performance improvement of 79% has been achieved for the decision tree model. We plan to use the indifferent and focused itemsets to improve performance of other prediction models.

References

1. Anwar, S., Rana, Z.A., Shamail, S., Awais, M.M.: Using association rules to identify similarities between software datasets. In: The 8th International Conference on the Quality of Information and Communications Technology (QUATIC), pp. 114–119. IEEE Computer Society, Lisbon (2012)

2. Baojun, M., Dejaeger, K., Vanthienen, J., Baesens, B.: Software defect prediction based on association rule classification. Open Access publications from Katholieke Universiteit Leuven urn:hdl:123456789/296322, Katholieke Universiteit Leuven (February 2011)
3. Boetticher, G., Menzies, T., Ostrand, T.: Promise repository of empirical software engineering data (2007)
4. Challagulla, V.U.B., Bastani, F.B., Paul, R.A.: Empirical assessment of machine learning based sofwtare defect prediction techniques. In: Proceedings of 10th Workshop on Object-Oriented Real-Time Dependable Systems (WORDS 2005), pp. 263–270. IEEE Computer Society, Washington, DC (2005)
5. Fenton, N.E., Neil, M.: A critique of software defect prediction models. IEEE Transactions on Software Engineering 25(5), 675–687 (1999)
6. Jiawei, H., Micheline, K.: Data Mining - Concepts and Techniques. Morgan Kaufmann (2002)
7. Kamei, Y., Monden, A., Morisaki, S., Matsumoto, K.-I.: A hybrid faulty module prediction using association rule mining and logistic regression analysis. In: Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2008, pp. 279–281. ACM, New York (2008)
8. Menzies, T., Greenwald, J., Frank, A.: Data mining static code attributes to learn defect predictors. Software Engineering, IEEE Transactions on 33(1), 2–13 (2007)
9. Rana, Z.A., Shamail, S., Awais, M.M.: Towards a generic model for software quality prediction. In: WoSQ 2008: Proceedings of the 6th International Workshop on Software Quality, pp. 35–40. ACM (May 2008)
10. Song, Q., Shepperd, M., Cartwright, M., Mair, C.: Software defect association mining and defect correction effort prediction. IEEE Transactions on Software Engineering 32(2), 69–82 (2006)
11. Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J.: The waikato environment for knowledge analysis, weka (2008)
12. Zafar, H., Rana, Z.A., Shamail, S., Awais, M.M.: Finding focused itemsets from software defect data. In: Proceedings of the 15th International Multitopic Conference (INMIC 2012). IEEE (2012)

Detection of Driving Fatigue Based on Grip Force on Steering Wheel with Wavelet Transformation and Support Vector Machine

Fan Li[1], Xiao-Wei Wang[1], and Bao-Liang Lu[1,2]

[1] Center for Brain-like Computing and Machine Intelligence
Department of Computer Science and Engineering
[2] MOE-Microsoft Key Lab. for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai 200240, P.R. China
bllu@sjtu.edu.cn

Abstract. This paper proposes an unobtrusive way to detect fatigue for drivers through grip forces on steering wheel. Simulated driving experiments are conducted in a refitted passenger car, during which grip forces of both hands are collected. Wavelet transformation is introduced to extract fatigue-related features from wavelet coefficients. We compare the performance of k-nearest neighbours, linear discriminant analysis, and support vector machine (SVM) on the task of discriminating drowsy and awake states. SVM with radial basis function reaches the best accuracy, 75% on average. The results show that variation in grip forces on steering wheel can be used to effectively detect drivers' fatigue.

Keywords: fatigue detection, grip force, wavelet transformation.

1 Introduction

The world vehicle population was reported to have surpassed the 1 billion-unit mark in 2010 (240 million in U.S. and 78 million in China) [1]. Automobile has been becoming the most important necessity for travel. However, accompanied with it is that more and more traffic accidents have happened. Driving fatigue has long been identified as one of the major causes of traffic accidents. It was founded that the crash risk was fourteen time higher for drives who had almost fallen behind the wheel [2]. According to National Highway Traffic and Safety Administration (NHTSA) report, driver fatigue and drowsiness causes 100,00 crashes annually, resulting in more than 40,000 injuries. If we can determine the onset of driving fatigue, such accidents can be avoided.

Most of the existing driving fatigue detection methods can be divided into three categories [3]:

1) Physical and physiological data of drivers are used to detect their driving fatigue. These include the Electroencephalography (EEG), Electrooculography (EOG) and eye patterns and head movement by video [4]. PERCLOS

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 141–148, 2013.

(PERcent eyelid CLOSure), put forward by Carnegie Melon Driving Research Center, is one of the most effective measure for driving fatigue detection [5].

2) Driving performance is indirectly assessed by raw rate, lateral position and longitude speed. Daimler Chrysler [6] has developed a detection algorithm that jointly analyzed these signals to detect drivers' fatigue.
3) Drivers in drowsy status would handle steering wheel and tread pedals (gas, brake and clutch) more slowly and improperly than in sober state [3,7,8]. Drivers would diminish grip force when falling into drowsy, even loosen the steering wheel fully, which could easily lead to accident. Thum Chia Chieh [7] proposed a statistical method to accumulate the logarithm of probability ratio of staying in drowsy or awake state. But the model was too simple to work well. Eskandarian and Mortazavi [8] trained an artificial neural network to detect drivers' fatigue, based on the hypothesis that under a drowsy state, the steering wheel movements become less precise and larger in amplitude. The proposed method had a big default that he did not fully consider time-order of steering wheel angle, which would lower detection accuracy.

Vigilance, the ability to maintain attention and alertness over prolonged periods of time, is an effect measurement of fatigue. In this paper, a simulated driving system was designed and subjects were asked to finish driving task in a real car. We collected drivers' grip force and response time to an audio signal which was used to measure drivers' vigilance while driving. We proposed an effective feature extraction method to extract features from time domain and wavelet coefficients through wavelet transformation. We compared the performance of three classifiers — support vector machine (SVM), linear discriminant analysis (LDA) and k-nearest neighbours (KNN), on the task of discriminating drowsy and awake states.

2 Experiment

2.1 Platform

Grip Force Detection

The setup for grip force detection is shown in Fig. 1. The wheel is fully covered by two pieces of force sensors to detect drivers' force of left and right hands. The force sensitive resistor (FSR) was available from Interlink Electronics, which is a very thin polymer thick film (PTR) device and will give little influence on driving. When force applied to its active surface, its resistance decreases. The resistance is converted to voltage through a simple resistance to voltage converter circuit. The output voltage V_{out} characterizes the force exerted on the wheel, and is collected by a commercial USB collection card with a multi-channel AD chip to PC for analysis. The sampling rate is 100 Hz for each hand's force.

Simulated Driving System

A car has been refitted for experiment. The car's driving operational devices (wheel and pedals) were replaced with Logitech's simulated controllers, whose

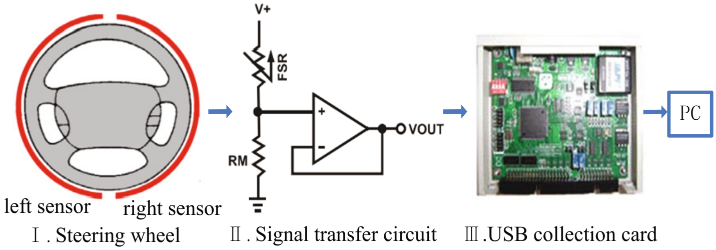


Fig. 1. Platform of detecting grip force of steering wheel

size and operation way are similar to real ones. We have developed a simulated driving software to imitate complex driving scenes with different weather, road conditions on a large LCD screen to cover subject's sight. The driving state parameters, such as driving speed, steering wheel angle, in curved or straight road and so on, will be recorded. Video cameras are installed to record subjects' face status and hand action, which can be aided to determine whether the subject is in a drowsy state.

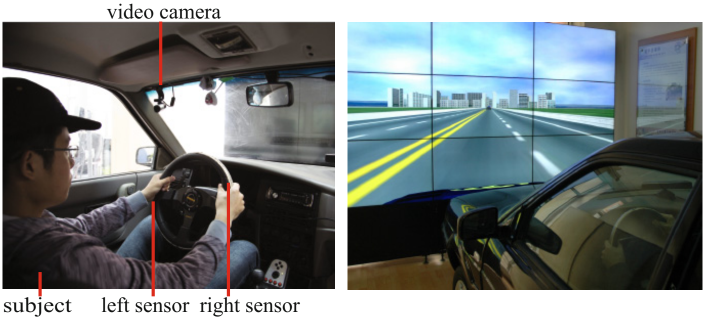


Fig. 2. Platform of simulated driving system

2.2 Procedure and Subjects

As shown in Fig. 2, subjects sit in the driver's seat to finish a two-hour-long driving task. They are asked to drive carefully to avoid crash with other cars. To measure subjects' vigilance, they need to do a periodical audio task which would not influence driving [9]. Fig. 3 shows the trial sequence. The trail period is 20 sec, and in the second half period, it will randomly play a two-second-long frog-croak sound. Once hear frog-croak, the subject would tread a special pedal with a resistance. When the pedal is trodden, its resistance value decreases, then through the same transfer circuit and USB collection card in Fig. 1, tread signal is recorded. During experiment, if the subject has fallen into sleep, he or she would be waken up to get a transitory sober status and restart driving. Four

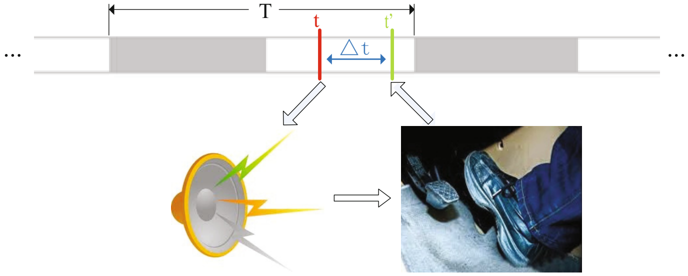


Fig. 3. The vigilance task. Trial period T is 20 s, t is the time of producing sound, t' is the tread time of subject, and Δt = t − t' is the response time.

healthy subjects of 19-25 years old have participated in this experiment for twice and the interval was one week. Experiments were carried out in a sound-proof room during 20:00 – 22:00 pm.

2.3 Data Collection

For each experiment, steering wheel angle signal is recorded by driving software at 100 Hz. Grip Force and tread signals are sampled by USB collection card at 100 Hz. And the time of sound played is recorded by PC, which is accurate to 1 millisecond. All these signals would be down sampled at 10 Hz and enough time stamps are made to synchronize them.

2.4 Vigilance Measurement

To train an effective fatigue detection method, reference vigilance indexes are necessary for supervised learning. In our experiment, the reaction time Δt_i of i-th trial can be used as reference vigilance level in that trial period, for there is an increase in average reaction time when the subject is becoming drowsy in vigilance task [10]. Because the cycle length of vigilance's fluctuation is usually longer than 4 min [11], to eliminate the variation of reaction time, we use a triangle weighted moving average (WMA) method to measure vigilance value in i-th trial period within a 2-min window:

$$vigilance_i = \frac{\sum_{j=i-3}^{i+3}(4 - |i-j|)\Delta t_j}{\sum_{j=i-3}^{i+3}(4 - |i-j|)} \tag{1}$$

Fig. 4 shows the vigilance curve after WMA for one experiment.

3 Method

3.1 Feature Extraction

The original signals we collect are two channels of time-varied grip force, through which it is difficult to detect the subject's temporal vigilance level. We can

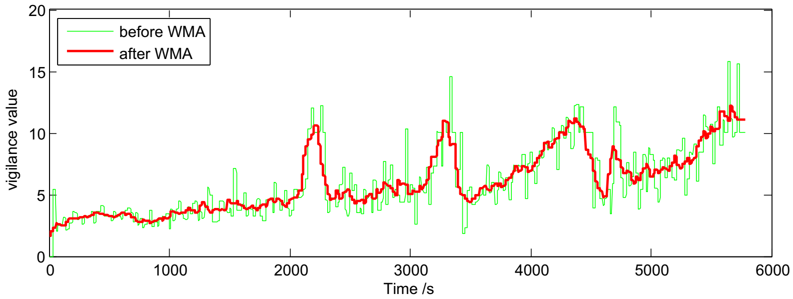


Fig. 4. Vigilance value represented by reaction time. The green curve is the original reaction time, and the red curve is the reaction time after WMA.

extract fatigue-related features in time domain and from wavelet coefficients of original signals through wavelet transformation.

The time-window length for one sample we choose is 25.6 sec. For the sample frequency is 10 Hz and interval is 0.1 sec, in a sample window, there are 256 data points of force, thus it is convenient to make 8-level wavelet transformation. And the moved step for sample windows is 6.4 sec, quarter of the length of a sample window.

Four kinds of statistical features are extracted in time domain. They are maximal value, minimal value, mean value and standard deviation, which can roughly characterize the range of force.

Because wavelet transformation is localized in both time and frequency, it can characterize the power spectrum in frequency domain. Haar wavelet function is chosen, so wavelet coefficients at i-th level, which characterize variation of signal in the corresponding time scale, can be calculated:

$$c_{ij} = r \cdot \left(\sum_{k=(j-1)\cdot 2^i+1}^{(j-1)\cdot 2^i+2^{i-1}} a_k - \sum_{(j-1)\cdot 2^i+2^{i-1}+1}^{j\cdot 2^i} a_k \right) \tag{2}$$

where $i = 1, 2, \ldots, 8$, $j = 1, 2, \ldots, \frac{l}{2^i}$, $l = 256$, and r is a positive normalization coefficient. The following features can be extracted from wavelet coefficients at i-th level to represent the time-frequency distribution of force signals:

1) Square sum of the wavelet coefficients, which represents the power of force signals in the corresponding frequency band.

$$p_i = \sum_j c_{ij}^2 \tag{3}$$

2) Ratio of positive coefficients in the whole wavelet coefficients.

$$pr_i = \frac{\sum_j 1 \cdot \{c_{ij} > 0\}}{\sum_j 1} \tag{4}$$

When grip force is increased, corresponding coefficient will be positive. It represents the time proportion of increasing grip force.

3) Ratio of sum of positive coefficients to sum of absolute value of negative coefficients, which characterizes the amount of increased force relative to decreased force. To calculate conveniently, logarithm value is used.

$$lpnr_i = \log_{10} \frac{\sum_j |c_{ij}| \cdot \{c_{ij} > 0\}}{\sum_j |c_{ij}| \cdot \{c_{ij} < 0\}} \tag{5}$$

Therefore, a total of 56 features can be extracted from two channels of force signals.

3.2 Classification

There exits difference in each subject's fatigue mode, so original labeled vigilance value was regularized firstly. By referring to recorded driving video, a specified threshold was chosen to determine whether the subject was stay in drowsy or awake status for each subject. In a two-hour-long experiment, there are about 700 cases.

We classified samples using three kinds of algorithms, k-nearest neighbors (KNN), linear discriminant analysis (LDA) classification and support vector machine (SVM). KNN is a non-parametric method for classifying the test case based on k most nearest training example. LDA is a linear classifier based on statistical characteristics of the training samples. SVM maps the input features into another feature space using kernel function, and iteratively approaches the optimal hyperplane with maximal margins. LIBSVM [12] was the tool to train SVM classifier we used, and the selected kernels were linear and radial basis function (RBF). At first, we employed these algorithms within one single experiment's samples. 3/4 samples were randomly chosen from drowsy set and awake set respectively as training samples, and the other as test samples. Then we attempted to use samples of the first experiment as training set to predict samples of the second experiment.

4 Results and Discussions

The classification performance of the aforementioned classifiers is listed in Table 1 and Table 2. Accuracies within one single experiment are higher than that between two experiments of one subject, due to strong correlation of force signals in one single experiment. The performance of LDA is close to SVMs, because of the fact that one sample only contains time sequence with 512 data points of grip force, a linear classifier with the 56 original features is sufficient to characterize the influence of subject's fatigue on grip forces. KNN is the worst one, and the accuracy of SVM-RBF is higher than SVM-linear and LDA, for it maps original features into more complex feature space. The performance of these classifiers is different among subjects.

Table 1. Classification accuracies of different classifiers within one single experiment

Classifier	Subject 1	Subject 2	Subject 3	Subject 4	Average
KNN	0.695	0.800	0.692	0.784	0.742
LDA	0.853	0.877	0.757	0.830	0.829
SVM-linear	0.848	**0.890**	0.757	0.824	0.830
SVM-RBF	**0.856**	0.862	**0.758**	**0.855**	**0.833**

Table 2. Classification accuracies of different classifiers between two experiments

Classifier	Subject 1	Subject 2	Subject 3	Subject 4	Average
KNN	0.652	0.686	0.636	0.677	0.663
LDA	0.720	**0.804**	0.670	0.680	0.719
SVM-linear	0.733	0.795	0.642	0.740	0.727
SVM-RBF	**0.751**	0.793	**0.674**	**0.783**	**0.750**

5 Conclusion

In this paper, we have proposed an unobtrusive drivers' fatigue detection method based on grip force on steering wheel. Wavelet transformation was used to extract fatigue-related features. By comparing performance of three kinds of classifiers, SVM with radial basis function reaches the best accuracy, but it needs more time for training model and detecting fatigue in practice. As a trade-off, SVM with linear kernel is preferred, for it is faster and owns a good classification accuracy. The experiment was conducted in a simulated environment, and in the future, we will conduct experiment in real driving environment to check the feasibility of the proposed method.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (Grant No. 61272248), the National Basic Research Program of China (Grant No. 2013CB329401), and Science and Technology Commission of Shanghai Municipality (Grant No. 13511500200).

References

1. Sousanis, J.: World vehicle population tops 1 billion units, Ward Auto World (August 15, 2011), `http://wardsauto.com/ar/world_vehicle_population_110815`
2. Lyznicki, J.M., Doege, T.C., Davis, R.M., Williams, M.A., et al.: Sleepiness, driving, and motor vehicle crashes. JAMA: Journal of the American Medical Association 279(23), 1908–1913 (1998)
3. Desai, A.V., Haque, M.A.: Vigilance monitoring for operator safety: A simulation study on highway driving. Journal of Safety Research 37(2), 139–147 (2006)
4. Du, R.-F., Liu, R.-J., Wu, T.-X., Lu, B.-L.: Online vigilance analysis combining video and electrooculography features. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012, Part V. LNCS, vol. 7667, pp. 447–454. Springer, Heidelberg (2012)

5. Grace, R., Byrne, V.E., Bierman, D.M., Legrand, J.-M., Gricourt, D., Davis, B., Staszewski, J.J., Carnahan, B.: A drowsy driver detection system for heavy vehicles. In: Proceedings of the 17th AIAA/IEEE/SAE Digital Avionics Systems Conference, DASC 1998, vol. 2, p. I36-1. IEEE (1998)
6. Renner, G., Mehring, S.: Lane departure and drowsiness – two major accident causes-one safety system. In: Mobility for Everyone. 4th World Congress on Intelligent Transport Systems, Berlin, October 21-24 (1997); paper No. 2264
7. Chieh, T.C., Mustafa, M.M., Hussain, A., Zahedi, E., Majlis, B.: Driver fatigue detection using steering grip force. In: Proceedings of the Student Conference on Research and Development, SCORED 2003, pp. 45–48. IEEE (2003)
8. Eskandarian, A., Mortazavi, A.: Evaluation of a smart algorithm for commercial vehicle driver drowsiness detection. In: IEEE Intelligent Vehicles Symposium, pp. 553–559 (2007)
9. Shi, L.-C., Lu, B.-L.: Eeg-based vigilance estimation using extreme learning machines. Neurocomputing 102, 135–143 (2013)
10. Smith, M.E., McEvoy, L.K., Gevins, A.: The impact of moderate sleep loss on neurophysiologic signals during working-memory task performance. Sleep 25(7), 784 (2002)
11. Makeig, S., Inlow, M.: Lapse in alertness: coherence of fluctuations in performance and eeg spectrum. Electroencephalography and Clinical Neurophysiology 86(1), 23–35 (1993)
12. Chang, C.-C., Lin, C.-J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), http://www.csie.ntu.edu.tw/~cjlin/libsvm

Three-Way Nonparametric Bayesian Clustering for Handwritten Digit Image Classification

Tomonari Masada[1] and Atsuhiro Takasu[2]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki 8528521, Japan
masada@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 1018430, Japan
takasu@nii.ac.jp

Abstract. This paper proposes a new approach for handwritten digit image classification using a nonparametric Bayesian probabilistic model, called multinomialized subset infinite relational model (MSIRM). MSIRM realizes a three-way clustering, i.e., a simultaneous clustering of digit images, pixel columns, and pixel rows, where the numbers of clusters are adjusted automatically with Chinese restaurant process (CRP). We obtain MSIRM as a modification of subset infinite relational model (SIRM) by Ishiguro et al. [4] While this modification is straightforward, our application of MSIRM to handwritten digit image classification leads to an impressive result. To represent a large number of training digit images in a compact form, we cluster the training images and then classify a test image to the class of the cluster most similar to the test image. By extending this line of thought, MSIRM clusters not only digit images but also pixel columns and pixel rows to obtain a more compact representation. With this three-way clustering, we achieved 2.95% and 5.38% test error rates for MNIST and USPS datasets, respectively.

Keywords: clustering, Bayesian nonparametrics, handwritten digit recognition.

1 Introduction

Handwritten digit recognition is an old problem. The Web site of well-known MNIST dataset[1] presents many evaluation results achieved by a great variety of methods. This paper proposes a new probabilistic approach widely different from those that can be found on the MNIST Web site. Our approach works like document-word co-clustering [2] in text mining. That is, we cluster not only digit images but also pixels. More precisely, we cluster digit images, pixel columns and pixel rows, simultaneously. In this manner, we conduct a *three-way* clustering over training digit images (cf. Fig. 1). We could not find any existing approaches similar to ours. Therefore, it can be said that our main contribution is to provide a completely new solution for handwritten digit recognition.

The advantages of our approach are as follows:

[1] http://yann.lecun.com/exdb/mnist/

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 149–156, 2013.

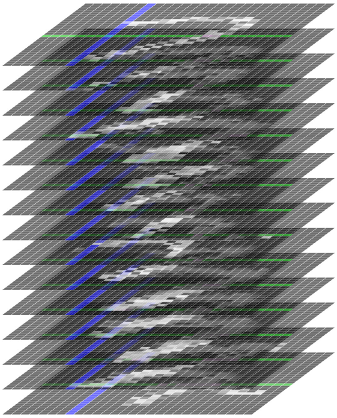


Fig. 1. This figure shows piled-up images of handwritten digit 7 from MNIST dataset. We cluster not only digit images but also pixel columns (blue) and pixel rows (green).

1. We adopt *bag-of-features* philosophy and conduct clustering based on feature frequencies. All features are discrete and have no intrinsic relation to one another. The features presented in this paper can be replaced with other features without modifying our approach.
2. Our approach is *not instance-based*. We represent training images in a compact form by clustering them. When we classify a test image, we calculate its similarity to each cluster and assign it to the class of the cluster giving the largest similarity. We use less memory than instance-based approaches.
3. We *automatically determine the number of clusters* with Chinese restaurant process (CRP) as in [4], which is a well-known technique of Bayesian nonparametrics. Consequently, we can reduce the number of free parameters.
4. We detect *irrelevant pixels*. Pixels in the peripheral part of digit images are often filled with the background color. Our approach detects such pixels as irrelevant and makes classification faster by ignoring irrelevant pixels.
5. We conduct a *three-way clustering*. That is, we cluster not only digit images but also pixel columns and pixel rows. By piling up all training images of the same class, we obtain a 3D rectangle as is depicted in Fig. 1. We conduct a three-way clustering on this 3D rectangle. Z-clustering, i.e., clustering in the vertical direction, corresponds to clustering of training images. X- and Y-clusterings correspond to clusterings of pixel columns and of pixel rows, respectively. We will show that this three-way clustering works.

While the method by Masada et al. [8] shares the first four advantages, the authors do not consider image clustering and gives no remarkable accuracies. In contrast, our approach gives accuracies close to those of previous works.

The rest of the paper is organized as follows. Section 2 reviews previous works. Section 3 describes the three approaches compared in the experiment. Section 4 presents evaluation results obtained for MNIST and USPS datasets. Section 5 concludes the paper with future work.

2 Previous Works

The Web site of MNIST dataset provides the following three approaches that can give remarkable classification accuracies: k-NN, SVMs, and neural networks, where we regard convolutional networks as a variant of neural networks. This section discusses how our approach is different from them.

Firstly, since k-NN is an instance-based approach, it is widely different from our approach. We conduct a clustering of training images and thus represent a large number of training images in a compact form. *Secondly*, SVMs are a discriminative approach where no probabilistic models of observed data are considered. In contrast, our approach uses a probabilistic model, called MSIRM, which gives a description of how observed data are generated. Therefore, our approach is a generative one and is widely different. *Thirdly*, neural networks are the approach giving the best classification accuracy ever published for MNIST dataset. One of the outstanding advantages of this approach is that an elaborated feature extraction is automatically realized with the multi-layered architecture of neural networks. This advantage leads to excellent accuracies. However, the work giving the best accuracy for MNIST dataset [1] contains no results for another famous dataset, USPS dataset[2]. Further, it is not clear how the authors could determine the highly complicated architecture of their neural networks, because the architecture is, in a sense, a free parameter. By inspecting several works using both datasets, not restricted to those adopting neural networks, we found that USPS tended to give an accuracy worse than MNIST. Halkias et al. [3] achieve 1.17% and 5.15% test error rates for MNIST and USPS, respectively. Maji et al. [7] achieve 0.56% and 2.7% error rates with SVMs. Keglevic et al. [5] describe USPS as “more challenging” than MNIST and give 0.76% and 2.84% error rates with SVMs under a setting different from [7]. Keysers et al. [6] give 1.0% and 2.2% error rates with an instance-based approach.

We will explain later that our approach achieves 2.95% and 5.38% test error rates for MNIST and USPS datasets, respectively. For MNIST dataset, we need to find more effective settings when compared with the existing results. However, our result for USPS dataset is close to that of [3]. While the main contribution of this paper is to provide a new approach for handwritten digit recognition, further improvements can follow through using a sophisticated feature extraction.

3 Compared Approaches

This section contains a description of our three-way clustering along with descriptions of the two compared methods. However, before giving the descriptions, we explain what kind of features we use in these three compared methods.

We extract image features as follows. 1) We apply the 3x3 pixels Gaussian blur filter. 2) At each pixel, we calculate brightness differences with respect to the upper, lower, left, and right pixels and encode each difference as a ternary value from $\{-, 0, +\}$ depending on its sign. This feature extraction gives $3^4 = 81$

[2] `http://www-i6.informatik.rwth-aachen.de/%7Ekeysers/usps.html`

different features. 3) We reduce the brightness of every pixel to a binary value from $\{0, 1\}$ depending on whether the brightness is smaller than half of the full brightness or not. We combine these two types of features and obtain $81 \times 2 = 162$ different features in total. Consequently, every pixel of digit images can be regarded as an occurrence of one among the 162 features. In this paper, we use these simple features, which can be sophisticated in the future.

We compared the following three approaches: one-way clustering, two-way clustering, and our three-way clustering. The first one is the baseline approach. We introduce some notations. Let N, I, and J denote the numbers of digit images, pixel columns, and pixel rows, respectively. For MNIST dataset, $N = 60,000$ and $I = J = 28$. For USPS dataset, $N = 7,291$ and $I = J = 16$. Let x_{nij} be an observed variable referring to the feature occurring at the pixel (i, j) in the nth training image, where i and j mean the ith column and the jth row, respectively. $\boldsymbol{X} = \{x_{nij}\}$ is our observed data for all compared methods.

The baseline method clusters training images and thus conducts a simple *one-way* clustering. Let z_n be a latent variable representing the cluster to which the nth training image belongs. The probability of training images can be written as $p(\boldsymbol{X}, \boldsymbol{z}) = p(z_1, \ldots, z_N) \prod_{n=1}^{N} p(\boldsymbol{x}_n | z_n)$. We run an MCMC inference [10] and iteratively update the values of $z_1, \ldots, z_N$. With respect to the cluster assignment probability $p(z_1, \ldots, z_N)$, we adopt Chinese restaurant process (CRP). Let K be the number of clusters. We obtain a probability $p(z_n = k | \boldsymbol{z}^{\setminus n}) \propto N_k^{\setminus n}$ for the existing clusters $k = 1, \ldots, K$, where $\setminus n$ means a removal of the nth image, and $N_k^{\setminus n}$ is the number of images belonging to the kth cluster after removing the nth image. For the new cluster, we obtain $p(z_n = K + 1 | \boldsymbol{z}^{\setminus n}) \propto \alpha_0$, where α_0 is the hyperparameter of CRP. We set $\alpha_0 = 1$, because other settings gave no significant differences. The observed data probability $\prod_n p(\boldsymbol{x}_n | z_n)$ given the latent variables $\boldsymbol{z}$ can be written as $\prod_n \prod_{i,j} \phi_{z_n i j x_{ij}}$, where ϕ_{kijw} denotes a probability that the wth feature occurs at the pixel (i, j) in the images belonging to the kth cluster. We apply a Dirichlet prior $\mathrm{Dir}(\boldsymbol{\beta})$ to all $\boldsymbol{\phi}_{kij}$s and integrate $\boldsymbol{\phi}_{kij}$s out. Therefore, the probability is rewritten as $\prod_k \prod_{i,j} \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(c_{kijw} + \beta_w)}{\Gamma(c_{kij} + \sum_w \beta_w)}$, where c_{kijw} is the number of occurrences of the wth feature at the pixel (i, j) of the images belonging to the kth cluster, and $c_{kij} \equiv \sum_w c_{kijw}$. Consequently, we obtain a probability for MCMC update of the cluster assignment of the nth image as $p(z_n = k | \boldsymbol{X}, \boldsymbol{z}^{\setminus n}) \propto N_k^{\setminus n} \times \prod_{i,j} \frac{\prod_w (c_{kijw}^{\setminus n} + \beta_w)}{c_{kij}^{\setminus n} + \sum_w \beta_w}$ for the existing kth cluster and as $p(z_n = K + 1 | \boldsymbol{X}, \boldsymbol{z}^{\setminus n}) \propto \alpha_0 \times \prod_{i,j} \frac{\prod_w \beta_w}{\sum_w \beta_w}$ for the new cluster.

We call the second compared method *two-way clustering*, because this method clusters not only digit images but also pixels. Let c_{klw} denote the number of occurrences of the wth feature at the pixels belonging to the lth pixel cluster in the digit images belonging to the kth image cluster. Let u_h be a latent variable representing the pixel cluster to which the hth pixel belongs. In our MCMC inference, we alternately update image cluster assignments $\boldsymbol{z}$ and pixel cluster assignments $\boldsymbol{u}$. We make $\setminus h$ mean a removal of the hth pixel and let $H_l^{\setminus h}$ denote the number of pixels belonging to the lth pixel cluster after removing the hth pixel. Let L be the

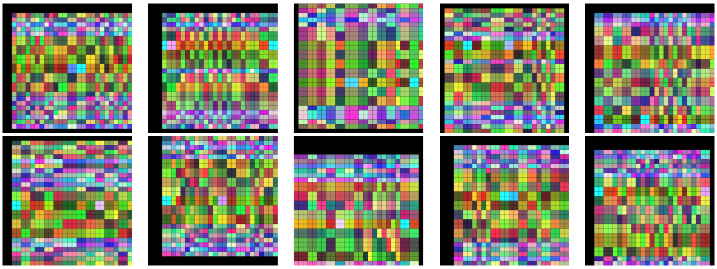


Fig. 2. From the top left panel to the bottom right panel, we give an example of clustering of pixel columns and pixel rows for from class "0" to "9" of MNIST dataset, respectively. Pixels belonging to the same pair of pixel column cluster and pixel row cluster are in the same color. Irrelevant pixel columns and rows are in black.

number of pixel clusters. While we skip details due to space limitation, we obtain a probability for MCMC update of the pixel cluster assignment of the hth pixel as $p(u_h = l|\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{u}^{\setminus h}) \propto H_l^{\setminus h} \times \prod_k \frac{\prod_w (c_{klw}^{\setminus h} + c_{khw} + \beta_w)}{c_{kl}^{\setminus h} + c_{kh} + \sum_w \beta_w} \frac{c_{kl}^{\setminus h} + \sum_w \beta_w}{\prod_w (c_{klw}^{\setminus h} + \beta_w)}$ for the existing lth pixel cluster and $p(u_h = L+1|\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{u}^{\setminus h}) \propto \alpha_1 \times \prod_k \frac{\prod_w (c_{khw} + \beta_w)}{c_{kh} + \sum_w \beta_w} \frac{\sum_w \beta_w}{\prod_w \beta_w}$ for the new cluster, where c_{khw} is the number of occurrences of the wth feature at the hth pixel of the images belonging to the kth image cluster, $c_{kh} \equiv \sum_w c_{khw}$, and α_1 is the hyperparameter of CRP for pixel clustering. α_1 is also set to 1.

The third compared method is our *three-way clustering* based on MSIRM we obtain by modifying SIRM [4]. In MSIRM, we have three types of clusters, i.e., image clusters, pixel column clusters, and pixel row clusters. Let u_i be a latent variable representing the pixel column cluster to which the ith pixel column belongs and v_j be a latent variable representing the pixel row cluster to which the jth pixel row belongs. Our MCMC inference alternately updates three types of latent variables, i.e., $\boldsymbol{z} = \{z_1, \ldots, z_N\}$, $\boldsymbol{u} = \{u_1, \ldots, u_I\}$, and $\boldsymbol{v} = \{v_1, \ldots, v_J\}$. As in SIRM, we can detect an irrelevant subset for each data domain by using special latent variables $\boldsymbol{r}_0$, $\boldsymbol{r}_1$, and $\boldsymbol{r}_2$ in MSIRM. $r_{0n} = 0$, $r_{1i} = 0$, and $r_{2j} = 0$ mean that the nth training image, the ith pixel column, and the jth pixel row are irrelevant, respectively. However, for both MNIST and USPS datasets, we obtained no irrelevant digit images. That is, $r_{0n} = 1$ for all n. This means that MSIRM regarded all training images as relevant for classification.

We present an example of three-way clustering results obtained for each of the ten classes of MNIST dataset in Fig. 2. The pixels filled with the same color belong to the same pair of pixel column cluster and pixel row cluster. Black pixel columns and rows are those that MSIRM detects as irrelevant. It can be considered that, since the pixels in these columns and rows are likely to be filled with the background color, MSIRM detects them as irrelevant. With respect to digit image clusters, we give an example of cluster size distributions obtained by MSIRM for all ten classes of MNIST dataset in Fig. 3, where we sort all clusters in the decreasing order of their sizes and plot them for each class.

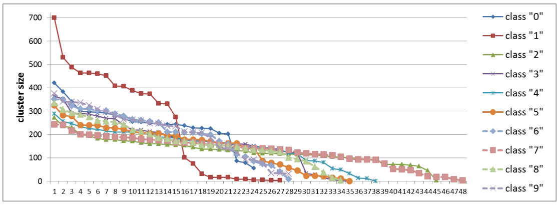


Fig. 3. Sizes of clusters given by MSIRM for MNIST in decreasing order

With respect to MSIRM, we can obtain probabilities for MCMC update of latent variables in a manner similar to SIRM by replacing binomial distributions with multinomial distributions and by increasing the number of domains from two to three. We refer readers to [4]. For example, we obtain a probability that the nth image is assigned to the kth existing image cluster as:

$$
\begin{aligned}
& p(z_n = k, r_{0n} = 1 | \boldsymbol{z}^{\setminus n}, \boldsymbol{r}_0^{\setminus n}, \boldsymbol{u}, \boldsymbol{r}_1, \boldsymbol{v}, \boldsymbol{r}_2, a_0, b_0, \boldsymbol{\gamma}, \boldsymbol{\beta}) \\
& \propto \frac{a_0 + R_0^{\setminus n}}{\Gamma\big(\sum_w (Q_w^{\setminus n} + q_{nw} + \gamma_w)\big)} \cdot \prod_w \left[\frac{\Gamma(Q_w^{\setminus n} + q_{nw} + \gamma_w)}{\Gamma(Q_w^{\setminus n} + \gamma_w)} \right] \cdot \frac{m_{0k}^{\setminus n}}{\alpha_0 + R_0^{\setminus n}} \\
& \cdot \prod_s \prod_t \left\{ \frac{\Gamma\big(\sum_w (c_{kstw}^{\setminus n} + \beta_w)\big)}{\Gamma\big(\sum_w (c_{kstw}^{\setminus n} + c_{nstw} + \beta_w)\big)} \prod_w \frac{\Gamma(c_{kstw}^{\setminus n} + c_{nstw} + \beta_w)}{\Gamma(c_{kstw}^{\setminus n} + \beta_w)} \right\} . \qquad (1)
\end{aligned}
$$

For all symbols in Eq. (1), the superscript $\setminus n$ means that we obtain the corresponding count after removing the nth digit image. a_0 and b_0 are the hyperparameters of the beta prior for the binomial distribution that determines the number of relevant and irrelevant images. R_0 is the number of relevant images. Q_w is the number of irrelevant pixels that are occurrences of the wth feature. q_{nw} is the number of irrelevant pixels that are occurrences of the wth feature and are placed in the nth image. γ_w is the hyperparameter of the Dirichlet prior for the multinomial distribution generating features at irrelevant pixels. m_{0k} is the number of images belonging to the kth image cluster. α_0 is the hyperparameter of CRP for digit image clustering. Note that the term $m_{0k}^{\setminus n}/(\alpha_0 + R_0^{\setminus n})$ comes from this CRP. c_{kstw} is the number of occurrences of the wth feature at the pixels simultaneously belonging to the sth column cluster and to the tth row cluster in the images from the kth image cluster. c_{nstw} is the number of occurrences of the wth feature at the pixels simultaneously belonging to the sth column cluster and to the tth row cluster in the nth digit image. In MCMC inference, we optimize Dirichlet hyperparameters with Minka's fixed point iteration [9].

We explain how we classify test images based on a result of clustering. In all of the three compared approaches, clustering is conducted separately on each of the ten classes, from class "0" to class "9". Consequently, the training images of each class are split into clusters. We calculate the similarity of a test image to each image cluster as the probability that the test image is generated as a member of the cluster. We calculate this probability based on latent variable values given by MCMC. For example, we consider the probability of a test image $\hat{\boldsymbol{x}}$ with respect

Table 1. The mean and standard deviation of error rates for 20 clustering results

	one-way clustering	two-way clustering	three-way clustering
MNIST	4.18±0.11%	3.66±0.30%	3.68±0.11%
USPS	6.63±0.20%	5.98±0.20%	6.33±0.30%

to the kth image cluster obtained by our three-way clustering conducted on the training images of class “7”. Let $\hat{x}_{ij}$ denote the feature occurring at the pixel (i, j) of the test image. When this pixel is relevant, its probability is written as $(c^{\text{“7”}}_{ku_iv_j\hat{x}_{ij}} + \beta_{\hat{x}_{ij}})/\sum_w(c^{\text{“7”}}_{ku_iv_jw} + \beta_w)$, where $c^{\text{“7”}}_{kstw}$ means c_{kstw} (cf. the previous paragraph) calculated for class “7”. The pixels in irrelevant columns and rows (cf. black pixels in Fig. 2) are regarded as generated with probability 1. This means that we ignore irrelevant pixels in calculation of test image probabilities. Therefore, we can accelerate test image classification. In this manner, we take advantage of the fact that MSIRM can detect irrelevant pixels. By multiplying the probabilities across all pixels in the test image, we obtain the probability that the test image belongs to the kth image cluster of class “7”. We calculate the probability of all test images with respect to all image clusters of all classes in this manner and classify them to the class of the cluster giving the largest probability, because the largest probability means the largest similarity.

4 Experiment

Our evaluation experiment compared the above three approaches, i.e., one-way, two-way, and our three way clusterings, on MNIST and USPS datasets. Our evaluation measure is test error rate, i.e., the percentage of misclassified test images. The numbers of test images are 10,000 and 2,007 for MNIST and USPS datasets, respectively. We conducted 300 iterations of MCMC inference of each approach after initializing latent variables randomly. We ran this MCMC 20 times and obtained 20 clustering results for each dataset by each approach.

We compare the three clustering approaches in two ways. *Firstly*, we give the mean and standard deviation of 20 test error rates corresponding to 20 clustering results (cf. Table 1). By using two-sided t-test at 5% significance level, we can say that both two-way and three-way clusterings give error rates smaller than the baseline. While three-way clustering gives no better result than two-way clustering, it accelerates classification, because we can ignore irrelevant pixels in probability calculation. *Secondly*, we select ten clustering results randomly from the 20 results obtained by each approach and merge the selected ten results to make an enlarged collection of image clusters. We repeat this ten times and make ten different enlarged collections of clusters. Then we use each of these ten enlarged collections for classifying test images. The number of clusters used for classification is ten-times larger than that in the first evaluation. Therefore, we can obtain a smaller error rate by using a ten-times larger space for storing cluster data. Table 2 contains the mean and standard deviation of ten test error rates corresponding to ten enlarged cluster collections. Our method shows an

Table 2. The mean and standard deviation of error rates for ten cluster collections

	one-way clustering	two-way clustering	three-way clustering
MNIST	3.41±0.05%	2.94±0.05%	2.95±0.06%
USPS	5.86±0.11%	5.54±0.16%	5.38±0.15%

advantage over two-way clustering for USPS and gives a comparable error rate for MNIST. This may be because our method gives more divergent clustering results when compared with two-way clustering. It can be concluded that, our method is better than one-way clustering and is comparable with two-way clustering, though our method can accelerate classification by ignoring irrelevant pixels.

5 Conclusion

We propose a completely new approach for handwritten digit recognition. Our method is a three-way clustering and is better than one-way clustering in test error rate as the experiment shows. While two-way clustering achieves almost the same error rate, our method can detect irrelevant pixels and thus can accelerate test image classification by ignoring irrelevant pixels. It is an important future work to improve the error rate further e.g. by using more sophisticated features.

References

1. Ciresan, D.C., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Proc. of CVPR 2012, 3642–3649 (2012)
2. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proc. of KDD 2001, pp. 269–274 (2001)
3. Halkias, X., Paris, S., Glotin, H.: Sparse penalty in deep belief networks: using the mixed norm constraint. In: Proc. of ICLR 2013 (2013)
4. Ishiguro, K., Ueda, N., Sawada, H.: Subset infinite relational models. JMLR W&CP 22, 547–555 (2012); AISTATS 2012
5. Keglevic, M., Sablatnig, R.: Digit recognition in handwritten weather records. In: Proc. of OAGM/AAPR Workshop (2013)
6. Keysers, D., Dahmen, J., Theiner, T., Ney, H.: Experiments with an extended tangent distance. In: Proc. of ICPR 2000, pp. 38–42 (2000)
7. Maji, S., Malik, J.: Fast and Accurate Digit Classification. Tech. Rep. No. UCB/EECS-2009-159 (2009)
8. Masada, T., Takasu, A.: Trimming prototypes of handwritten digit images with subset infinite relational model. In: Proc. of MUE 2013, pp. 129–134 (2013)
9. Minka, T.P.: Estimating a Dirichlet distribution (2000), `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`
10. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Statist. 9(2), 249–265 (2000)

A Universal Visual Dictionary Learned from Natural Scenes for Recognition

Li Ding and Jinhua Xu

Department of Computer Science and Technology
East China Normal University
Shanghai, China
jhxu@cs.ecnu.edu.cn

Abstract. Inspired by the efficient coding hypothesis and simple-to-complex cell hierarchy of the visual system, we study a universal visual dictionary learned from natural scenes using sparse coding for recognition. The vocabularies are similar to V1 simple cells receptive fields. Max pooling is done in a local region (”block”) so that the features are translation invariant, which is the function of complex cells. Macro-features of a grid of overlapping spatial blocks are built and fed to a linear SVM classifier for recognition. We have tested the learned universal visual dictionary on different recognition tasks and demonstrated the effectiveness of the model.

Keywords: sparse coding, dictionary learning, object recognition.

1 Introduction

A fundamental function of the visual system is to encode the building blocks of natural scenes-edges, textures and shapes-that subserve visual tasks such as object recognition and scene understanding.Many algorithms have been proposed to model the cortical image representation of human visual system, for example, the feedforward HMAX model [1] and deep belief networks[2].

Sparse coding was first introduced to compute a sparse representation of the natural stimuli data [3]. The dictionary learned from natural scene using sparse coding is similar to the receptive fields of V1 simple cells. Inspired by the work from neuroscience community, there has been an increasing interest on sparse image representation. Dictionaries for sparse coding have been mostly learned from training images in the problem domain in an unsupervised way [4,5]. Some work studied discriminative dictionary learning for recognition [6,7,8,9,10]. Back-propagation was applied to learn the dictionary[8,9,10]. It is not clear if the problem domain dictionary learned in a supervised or unsupervised way is relevant or necessary for object recognition.

In this paper, we evaluate the effectiveness of a universal dictionary. We first apply the sparse coding algorithm on natural images to learn the universal visual words that mimic the tuning properties of V1 simple cells. The dictionary is generic and universal in the sense that it is learned from natural scenes in

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 157–165, 2013.

an unsupervised way and can support the recognition of many different object categories. Then we perform a max-pooling operation in a larger region. The result of the pooling over positions is that the features become insensitive to the location of the stimulus, which is a hallmark of cortical complex cells. The proposed model, which is referred to ScMAX, is still a single-layer model, and the cortex hierarchy is not implemented yet. Therefore, to encode the spatial relationship, features of a grid of densely and uniformly spaced blocks are concatenated to a macro-feature, which is used to represent an object and fed to a classifier for recognition. We tested the proposed approach on different recognition tasks, and demonstrated that the universal dictionary is very efficient and powerful for different object representation.

2 Related Work

Some biologically motivated models have been proposed for object recognition. A feedforward model with features inspired visual cortex was proposed in [1]. It suggested that a task-independent, unsupervised, developmental-like learning stage may exist in the ventral stream to generate a large generic dictionary of shape-tuned units with various degrees of selectivity and invariance from V1 to IT, consistent with recent anatomical and physiological data. In HMAX model [1], the filters of the first layer are pre-defined Garbor filters with four orientations, while in this paper we use a universal visual dictionary learned from natural images using sparse coding, the resulting filters have various locations, orientations and spatial bandwidths, therefore should be more efficient and powerful.

The idea of universal visual feature was also proposed in [11], which suggests that all visual stimuli share some characteristic in common such that knowledge obtained from one set of stimuli can be applied to a completely different set of visual stimuli. It was observed that the common property shared by the appearance of human faces and the backyard view is the local statistical structure. The difference between our approach with Shan and Contrell [11] is neither block partition nor pooling was done in their approach.

In [5], a self-taught learning algorithm was proposed that uses sparse coding to construct higher-level features using the unlabeled data in classification tasks. The dictionaries learned from unlabeled data were applied to represent labeled data and the classification performance was improved, for example a dictionary learned from hand-written digits was used to represent hand-written characters. In this paper, we extended their work and evaluated a generic and universal dictionary for different recognition tasks.

3 Method

3.1 Sparse Coding

The goal of sparse coding is to represent input vectors approximately as a weighted linear combination of a small number of (unknown) ”basis vectors” or vocabularies. These basis vectors thus capture high-level patterns in the input data.

Let X be a set of image patches, i.e. $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, here N is the number of patches,and $\mathbf{x}_i \in R^n$(n is the number of pixels in a patch). Sparse coding can be formulated as the following optimization problem[3].

$$\min_{\mathbf{D},\mathbf{S}} \sum_{i=1}^{N} \{\|\mathbf{x}_i - \mathbf{D}\mathbf{s}_i\|_2^2 + \gamma \|\mathbf{s}_i\|_1\} \tag{1}$$

Here $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_v]$ is the learned basis or dictionary, and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N]$ is the sparse code of the patches. The basis set can be over-complete ($v > n$), and can thus capture a large number of patterns in the input data. γ is a regularization parameter which control the tradeoff between reconstruction error and sparsity.

Sparse coding has a training phase and a coding phase. In the training phase, Eq. (1) is solved with respect to $\mathbf{D}$ and $\mathbf{S}$; In the coding phase, for each image patch, the sparse code is obtained by optimizing Eq. (2) with respect to $\mathbf{s}$ only:

$$\min_{\mathbf{s}} \{\|\mathbf{x} - \mathbf{D}\mathbf{s}\|_2^2 + \gamma \|\mathbf{s}\|_1\} \tag{2}$$

We applied the sparse coding algorithm on ten 512x512 natural images available from Olshausens homepage [3]. We normalized each image to have zero mean and unit variance, then extracted 161290 8x8 image patches and subtracted the local mean from each image patch. The learned 128 basis (i.e. columns of $\mathbf{D}$) are shown in Fig.1(top)which are the universal dictionary used in the experiments in section 4. It should be pointed out that dictionary learning has been extensively studied in recent years. However, most of the dictionaries were learned from the training example, thus are problem dependent [4,6,10], as in Fig.1(bottom). In this paper, the dictionary was learned from general natural scenes.

3.2 Max Pooling

Max pooling procedure is well established by biophysical evidence in visual cortex (V1) and is used in many algorithms [1,4] to extract invariant features.

In our model, an image is partitioned into BxB blocks. The block size is w*h pixels. In each block, we extracted patches with PxP pixels, and then calculated the sparse code $\mathbf{s}^p = [s_1^p, ..., s_v^p]^T$ for each patch using (2). The local descriptor of a block $\mathbf{s}^b = [s_1^b, ..., s_v^b]^T$ was obtained by max pooling over all patches inside the block. The superscripts p and b are for patch and block respectively.

$$s_j^b = \max_p s_j^p, j = 1, ..., v \tag{3}$$

After max pooling over positions, we obtained the features for all blocks, which were then concatenated to form a large feature vector $\mathbf{z}$ in (4). This feature vector will be fed to a linear SVM classifier for recognition.

$$\mathbf{z} = [\mathbf{s}^1, \mathbf{s}^2, ..., \mathbf{s}^{B*B}] \tag{4}$$

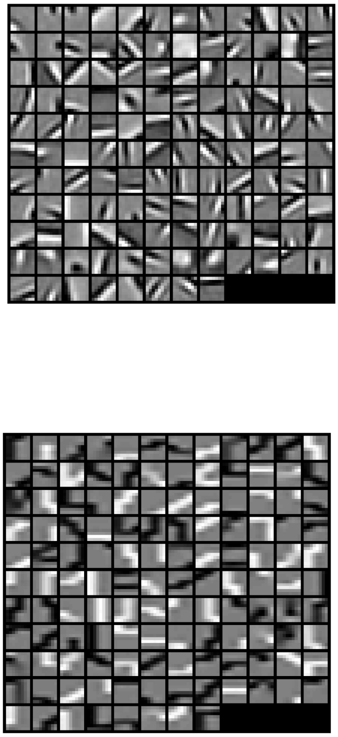


Fig. 1. Dictionaries learned from natural images (top) and from English characters images (bottom)

4 Experiments and Results

In this section, we use the universal visual dictionary learned from natural scene for different recognition tasks, including handwritten digits, handwritten English characters and faces.

4.1 USPS Dataset

The USPS dataset has 7291 training images and 2007 test images of size 16x16. The images of this dataset were partitioned to 2x2 blocks. We extracted patches with 8x8 pixels in each block. To demonstrate the advantage of the proposed approach, we compared our approach with a baseline approach, that is, the patch size (16x16) is the same as the image size. No pooling was performed in this case, and the sparse codes of the patch were fed directly to a classifier.

First, we tested different block sizes B= 8, 10, or 12. For B=8, the four blocks are non-overlapping, and there is only one patch in each block, therefore no pooling is needed; For B= 10 or 12, there are 4 or 8 pixels shared by two neighbor blocks. The results are shown in Table 1. It can be seen that the accuracies of block size B=10 or 12 are always better than that of B=8. This is because larger block is more insensitive to the shift variance due to the pooling operation.

Next, we tested different dictionary size v=64,128, or 256. As shown in Table 1, in the baseline approach, no pooling is performed, the accuracy for a larger dictionary size is better. However, in the proposed ScMAX, a larger dictionary size does not always mean a higher accuracy. This is contrast to the findings in [10]. It may be because the pooling operation in our approach compensates the weakness of a smaller dictionary size. Furthermore, the proposed approach performs better than the baseline approach for all dictionary sizes. The efficiency of the block structure and max pooling is verified.

Table 1. Recognition accuracy for USPS for different dictionary sizes and block sizes

Block size(B)	Dictionary size(V)		
	64	128	256
8	95.01	96.51	96.46
10	97.86	98.45	98.60
12	97.41	98.80	98.71
16(Baseline)	91.33	94.37	95.32

Finally, we compared our results with the state-of-the-art approaches in the literature. The best results of these approaches were shown in Table 2. Our error rate (ScMAX) is significantly better than other approaches. It should be noted in Mairal et al[10], no block partition and pooling was performed, the results in Table 2 for unsupervised and supervised approach in [10] were both for dictionary size 300, our result is for dictionary size 128. Their unsupervised approach is same as our baseline, except that they used a problem domain dictionary. In their supervised approach, the dictionary was not only in problem domain, but updated with back-propagation for recognition. However, our result with a smaller(128 versus 300) and universal dictionary is much better.

Table 2. Error rate for USPS for different approaches

approaches	Mairal et al [10] (unsupervised)	Mairal et al [10] (supervised)	ScMAX
Error rate	4.58	2.84	1.20

4.2 MNIST Dataset

The MNIST dataset has 60000 training examples and 10000 test examples. The digits have been size-normalized and centered in 28x28 images. We scaled the images to 24x24 pixels.Each image was partitioned to 3x3 blocks, and all patches of 8x8 pixels were extracted in each block. We tested three different block sizes,

Table 3. Recognition accuracy for MNIST for different block sizes

Block size (B)	8	10	12
Accuracy	95.64	98.80	99.21

B=8,10, or 12, and the dictionary size is 128. The recognition accuracy is shown in Table 3. The best result (99.21%) was obtained with block size 12.

We compared our results with some approaches in the literature. The best results of the approaches were shown in Table 4. Our error rate is slightly larger than the supervised dictionary learning in Mairal et al [10], but is smaller than the unsupervised dictionary learning in Mairal et al [10]. From the MNIST dataset website, the best accuracy on the dataset so far was obtained by the model in [12], which is a committee of 35 convolutional nets, much more complicated than the proposed ScMAX model. The approach in [4] is similar to ours, except that they used a problem domain dictionary and SVM classifier with Gaussian kernel. Their result was slightly better than ours because of the large training set (6000 training images per class).

Table 4. Error rate (%) for MNIST for different approaches

Approaches	Mairal et al [10] (unsupervised)	Mairal et al [10] (supervised)	Labusch et al [4]	ScMAX	Ciresan et al[12]
ErrorRate	2.36	0.54	0.59	0.79	0.23

4.3 Hand-Written Character Dataset

We also tested the proposed model on hand-written character dataset downloaded from (http://ai.stanford.edu/ btaskar/ocr/). The dataset has 52152 English characters. The image size is 16x8 pixels. We padded and scaled the images to size 24x24, as in [5,8]. Each image was partitioned into 3x3 blocks. Block size is 12x12, patch size is 8, and the dictionary size is 128. We randomly selected M images as the training set, the rest as the test set. The average accuracies of 50 runs for different M were shown in Table 5.

In order to compare the universal dictionary with the specific dictionary learned from training images, we tested a baseline approach with a problem domain dictionary, in which all parameters were same as in ScMAX. The dictionary learned from English character images is shown in Fig.1(bottom). Compared with the universal dictionary in Fig.1(top), it can be seen the dictionary atoms are for the specific problem, and cannot be used for other problems. The recognition results of the baseline are also in Table 5. It can be seen that the results of the universal dictionary are better when the training data set is less than 5000 and comparable with larger training sets. This conclusion is consistent with the self-taught learning in [5].

In [5], a basis learned by L1-regularized sparse coding on handwritten digits was shown to improve classification performance when used for handwritten character recognition with small training data sets. As shown in Table 5, our results from the universal dictionary are better with all training data sets than the results in [5].

In Bradley et al [8], a differentiable smooth KL prior for sparse coding was proposed to improve the prediction performance over L1-prior, and supervised dictionary learning through back-propagation further improved the performance. As shown in Table 5, their results are better than our ScMAX due to the KL prior and supervised learning.

Table 5. Recognition accuracy for hand-written character dataset

Training	Sparse coding [5]	L1[8]	KL[8]	KL+backprop[8]	BaseLine	ScMAX
100	39.7	44.0	49.4	50.7	37.2	42.8
500	58.5	63.7	69.2	69.9	59.3	62.1
1000	65.3	69.5	75.0	76.4	66.8	68.5
5000	-	78.9	82.5	84.2	78.6	78.8
20000	-	83.3	86.0	89.1	84.6	83.4

4.4 Face Recognition

We also evaluate the algorithms performance on face database CMU PIE[13]. The database consists of 41,368 images of 68 people, each person under 13 poses, 43 different illumination conditions and with 4 different expressions. In [9,14], a subset of the dataset was used, which contained only five near frontal poses (C05, C07, C09, C27,C29) under all illuminations. Therefore, there are 170 images for each individual. We use the same subset for fair comparison.

The size of each cropped face image is 64*64 pixels. We rescaled the images to size 24x24. Each image was partitioned into 3x3 blocks. Block size is 10x12, patch size 8x8, and the dictionary size 128. A random subset of images per person was selected as the training set and the rest of the database is considered as the testing set. We run 10 times for each case, the average training errors were shown in Table 6. We also run a baseline approach, in which the dictionary was learned from the face images. It can be seen that the universal dictionary performs better when the training set is small (≤ 70).

We compare our approach with previous model in Table 6. Our results are better than in [14], but not as good as in [9]. It is because supervised dictionary learning was used in [9], therefore, the dictionary was adapted to the specific recognition through back propagation, which is top down processing in the brain. Our model is a bottom-up processing.

Table 6. Recognition error rate for CMU PIE face dataset

Training	30	50	70	90	130
S-LDA[14]	3.6	2.5	2.1	1.8	1.6
S-SC[9]	0.49	0.15	0.12	0.037	0
Baseline	5.96	3.28	2.11	1.39	0.82
ScMAX	4.89	3.01	2.06	1.69	1.18

5 Conclusion

In this paper, a biologically motivated model was proposed for object recognition. Sparse coding was applied to natural images to learn a universal visual dictionary, which replicate the tuning properties of V1 simple cells and can be used to represent different kinds of objects, for example the hand-written digits, characters and faces. Max-pooling was performed inside a local region to make the feature position invariant, which mimics the function of complex cells. The model was tested on benchmark datasets using the learned universal visual dictionary, and it was found that the universal dictionary is more efficient than the dictionary in the problem domain when the training set is small.

In the future work, we will extend the simple-complex cell operations to the higher levels of the visual cortex hierarchy, and test the model on more complex objects, such as the ones in VOC PASCAL datasets.

Acknowledgment. This work is supported by the National Natural Science Foundation of China under Project 61175116, and Shanghai Knowledge Service Platform for Trustworthy Internet of Things (No. ZF1213).

References

1. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 994–1000 (2005)
2. Bengio, Y.: Learning deep architectures for ai. Foundations and Trends® in Machine Learning 2, 1–127 (2009)
3. Olshausen, B.A., et al.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
4. Labusch, K., Barth, E., Martinetz, T.: Simple method for high-performance digit recognition based on sparse coding. IEEE Transactions on Neural Networks 19, 1985–1989 (2008)
5. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning, pp. 759–766 (2007)
6. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Supervised dictionary learning. arXiv preprint arXiv:0809.3083 (2008)
7. Huang, K., Aviyente, S.: Sparse representation for signal classification. In: Advances in Neural Information Processing Systems (NIPS), pp. 609–616 (2006)

8. Bradley, D.M., Bagnell, J.A.: Differential sparse coding. In: Advances in Neural Information Processing Systems (NIPS) (2008)
9. Yang, J., Yu, K., Huang, T.: Supervised translation-invariant sparse coding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3517–3524 (2010)
10. Mairal, J., Bach, F., Ponce, J.: Task-driven dictionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 791–804 (2012)
11. Shan, H., Cottrell, G.W.: Looking around the backyard helps to recognize faces and digits. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
12. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3642–3649 (2012)
13. Sim, T., Baker, S., Bsat, M.: The cmu pose, illumination, and expression (pie) database. In: IEEE Conference on Automatic Face and Gesture Recognition, pp. 46–51 (2002)
14. Cai, D., He, X., Han, J.: Semi-supervised discriminant analysis. In: International Conference on Computer Vision (ICCV), pp. 1–7 (2007)

Affective Abstract Image Classification and Retrieval Using Multiple Kernel Learning

He Zhang, Zhirong Yang, Mehmet Gönen, Markus Koskela,
Jorma Laaksonen, Timo Honkela, and Erkki Oja

Department of Information and Computer Science
Aalto University School of Science, Espoo, Finland

{he.zhang,zhirong.yang,mehmet.gonen,markus.koskela,
jorma.laaksonen,timo.honkela,erkki.oja}@aalto.fi

Abstract. Emotional semantic image retrieval systems aim at incorporating the user's affective states for responding adequately to the user's interests. One challenge is to select features specific to image affect detection. Another challenge is to build effective learning models or classifiers to bridge the so-called "affective gap". In this work, we study the affective classification and retrieval of abstract images by applying multiple kernel learning framework. An image can be represented by different feature spaces and multiple kernel learning can utilize all these feature representations simultaneously (i.e., multiview learning), such that it jointly learns the feature representation weights and corresponding classifier in an intelligent manner. Our experimental results on two abstract image datasets demonstrate the advantage of the multiple kernel learning framework for image affect detection in terms of feature selection, classification performance, and interpretation.

Keywords: Image affect, multiple kernel learning, group lasso, low-level image features, image classification and retrieval.

1 Introduction

Multimedia contents such as audio, image, and video contain information that can trigger people's affective feelings or emotions. Such information can be used by search engines for better modeling the user's preferences. Affective image classification and retrieval has attracted increasing research attention in recent years, due to the rapid expansion of the digital visual libraries on the Web. While most of the current content-based image retrieval (CBIR) systems [6] are designed for recognizing objects and scenes such as plants, animals, outdoor places etc., an emotional semantic image retrieval (ESIR) system [17] aims at incorporating the user's affective states to enable queries like "beautiful flowers", "cute dogs", "exciting games", etc.

Though emotions are highly subjective human factors, still they have certain stability and generality across different people and cultures [12]. As an example, Figure 1 shows four pictures taken from an abstract art image collection [19]. The

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 166–175, 2013.

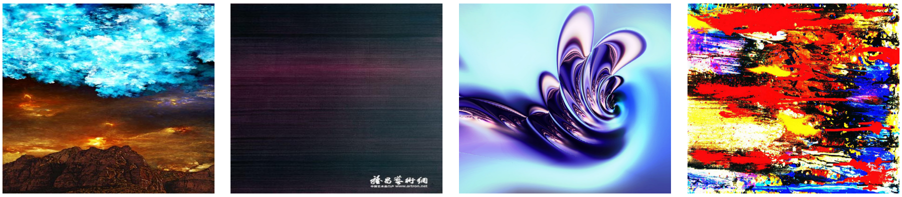

(a) Exciting (b) Boring (c) Relaxing (d) Irritating

Fig. 1. Example images from the abstract art image data set [19] with the ground truth labels of Exciting, Boring, Relaxing, and Irritating

ground truth labels are determined by the emotion that has received the most votes from people. Intuitively, the “Exciting” and “Relaxing” pictures usually make people feel pleasant or evoke a positive feeling, whereas the “Boring” and “Irritating” pictures may evoke a negative feeling to the viewer.

In analogy to the concept of “semantic gap” that implies the limitations of image content description, the “affective gap” can be defined as “the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal” [8]. Among the challenges from image affect detection, one is to select suitable image features to reflect people’s affective states, and another one is to build effective learning models or classifiers to bridge the “affective gap”.

Many works (e.g., [5,11]) have focused on designing features specific to image affect detection, while others (e.g., [14,19]) simply utilized the traditional low-level color, shape, and texture features. Concerning the classifiers, support vector machines (SVM) [4] have been adopted in most of the works. However, one usually has to spend much time and effort in picking up the most suitable feature representation that can best reflect the viewer’s emotions. For example, the authors in [14,19] utilized Fisher score to first rank and then select the most descriptive features, without considering the classifier at all. The authors in [5,11] picked each feature one by one with respectively an SVM and a naive Bayes classifier as the base learner to boost the performance, which requires explicit cross-validation steps for selecting features while optimizing the classifier parameters, and thus suffers from heavy computational complexities.

An image can be represented by different feature spaces. Multiple kernel learning (MKL) [2] can utilize all these feature representations simultaneously, such that it jointly learns the feature representation weights and the corresponding classifier for selecting automatically the most suitable feature representation or a combination of them. This can improve the classification performance and makes the interpretation of the results straightforward. MKL has earlier been applied for object detection in [16], and we are the first to introduce it into image affect detection. Our experimental results demonstrate the advantages of the MKL framework in affective classification and retrieval of abstract images.

Section 2 introduces the image features used in this paper. Section 3 introduces the MKL framework and an efficient algorithm that implements MKL. In Section 4, the experimental results on affective abstract image classification and retrieval are reported. Finally, the conclusions and future work are presented in Section 5.

2 Image Features

We have utilized a set of ten generic low-level color, shape, and texture features to represent each image. Table 1 gives a summary of these features. The features are extracted both globally and locally. For local features, a five-zone tiling mask is employed, where the image area is divided into four tiles by the two diagonals of the image, on top of which a circular center tile is overlaid [15]. All the features are extracted using the PicSOM system [10].

Table 1. The set of low-level image features used

Index	Feature	Type	Zoning	Dims.
F1	Scalable Color	Color	Global	256
F2	Dominant Color	Color	Global	6
F3	Color Layout	Color	8×8	12
F4	5Zone-Color	Color	5	15
F5	5Zone-Colm	Color	5	45
F6	Edge Histogram	Shape	4×4	80
F7	Edge Fourier	Shape	Global	128
F8	5Zone-Edgehist	Shape	5	20
F9	5Zone-Edgecoocc	Shape	5	80
F10	5Zone-Texture	Texture	5	40

Four of the features are standard MPEG-7 descriptors: Scalable Color, Dominant Color, Color Layout, and Edge Histogram. 5Zone-Color is defined as the average RGB values of all the pixels within the zone. 5Zone-Colm denotes the three central moments of HSV color distribution. Edge Fourier is calculated as the magnitude of the 16 × 16 FFT of Sobel edge image. 5Zone-Edgehist is the histogram of four Sobel edge directions. 5Zone-Edgecoocc is the co-occurrence matrix of four Sobel edge directions. Finally, 5Zone-Texture is defined as the histogram of the relative brightness of the neighboring pixels. More information about the features can be found in [15].

3 Multiple Kernel Learning

We can represent an image with different feature representations or views. However, the most suitable representation for a given task is generally not known a priori. Instead of using a single representation (i.e., single-view learning), we can also make use of different representations simultaneously (i.e., multiview learning). Multiview learning with kernel-based methods is known as multiple kernel learning, which is a principled way of combining kernels calculated on different views to obtain a better prediction performance than single-view learning methods (see [7] for an extensive survey). In addition, MKL can learn the feature representation weights by itself according to the data and task at hand during the training stage without an explicit feature selection step, which makes the interpretation easy and straightforward.

Among the MKL algorithms, we use the group Lasso MKL [1] as our learning framework in that it is simple and efficient [18]. Both studies [1,18] have formulated an alternating optimization method that solves an SVM at each iteration and updates the kernel or feature representation weights η_m as follows:

$$\eta_m = \frac{\|w_m\|_2^{\frac{2}{p+1}}}{\left(\sum_{h=1}^{P} \|w_h\|_2^{\frac{2p}{p+1}}\right)^{\frac{1}{p}}} \tag{1}$$

where $\|w_m\|_2^2 = \eta_m^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k_m(x_i^m, x_j^m)$ is from the duality conditions. $k_m(-,-)$ denotes the kernel function calculated on the mth feature representation. P is the number of kernels or feature representations ($P = 10$ in our case), and p is chosen to be 1 so that $\sum_{m=1}^{P} \eta_m = 1$.

After updating the kernel weights in equation (1), the algorithm then solves a classical SVM problem by maximizing SVM dual formulation with the combined kernel $k = \sum_{m=1}^{P} \eta_m k_m$ as follows:

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{2}$$

subject to the constraints: $0 \leq \alpha_i \leq C$ for all $i = 1, ..., N$, and $\sum_{i=1}^{N} \alpha_i y_i = 0$, where C is the regularization parameter and y_i is the label (± 1) of training sample x_i. The two steps alternate until convergence.

4 Experiments

In this section, we present the experimental results using the MKL framework in the classification and retrieval of abstract images. We implemented the group Lasso MKL in MATLAB and took 20 alternating iterations for inference. We chose the LIBSVM [3] package for solving the classical SVM problem. For the group Lasso MKL, We set $C = 1$ and calculated the standard Gaussian kernel on each feature representation separately with the kernel width $s = 2\sqrt{D_m}$, where D_m is the dimensionality of corresponding feature representation. Therefore, no cross-validation steps are needed for learning the feature representation weights or the parameters of SVM classifier in group Lasso MKL.

4.1 Datasets

We have chosen abstract art images as our learning target instead of the photographic images since the latter contain contextual information that may affect the viewer's emotional assessment, which in turn would bias the learning results. Two abstract image datasets have been used in the experiments, `Abstract100`[1] [19] and `Abstract280`[2] [11].

The `Abstract100` dataset contains 100 images of abstract art paintings with different sizes and qualities through Google image search. These paintings were originally created by artists with various origins and periods. Each image has been evaluated by 20 college students (10 females and 10 males) including Asians and Europeans for two descriptive/adjective pairs "Exciting vs. Boring" and "Relaxing vs. Irritating" from the ratings of $\{-2, -1, 0, 1, 2\}$.

The `Abstract280` dataset contains 280 abstract art images that were peer-rated from a Web survey. Each image was labeled as the single emotion that had received the most votes from the eight affective categories: Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sad(ness). The 280 images were rated by nearly 230 people, where each image was rated about 14 times.

4.2 Affective Abstract Image Classification

Experimental Setup. We use only the `Abstract100` dataset in this task, as SVM is optimized for binary classification problems. To obtain the ground truth labels for the classifier, we adopt a heuristic thresholding strategy: the image samples with ratings ≥ 0 in each descriptive pair are treated as the positive class, whereas those with ratings < 0 are treated as the negative class. For example, if an image receives an average rating of $(0.2, 1.5)$, then it is thresholded as $(+1, +1)$, which can be interpreted as both "Exciting" and "Relaxing". This results in roughly equal numbers of positive and negative samples. For training and testing, we use 5-fold cross-validation and calculate the average classification accuracy for each adjective pair. For comparisons, SVM_all uses the concatenation of all the 10 feature representations of an image as a single input, while SVM_best uses each of the 10 feature representations individually (as in [5,11]) and reports the one that has obtained the highest accuracy. Note that methods in both papers [5,11] require explicit cross-validation steps to select features and to optimize parameters (C and s), whereas no cross-validation procedures are involved in learning the adopted group Lasso MKL. The baseline result is calculated as the proportion of the majority class in each case.

Results. Figure 2 shows the average feature representation weights (i.e., kernel weights) in the range $[0, 1]$ based on 5-fold cross-validation using group Lasso MKL algorithm. We clearly see that, for the "Exciting vs. Boring" pair, Scalable Color (F1) ranks first, followed by Zone5-Color (F4), Edge-Histogram (F6), and

[1] An updated version: `http://research.ics.aalto.fi/cbir/abstract100`

[2] `http://www.imageemotion.org`

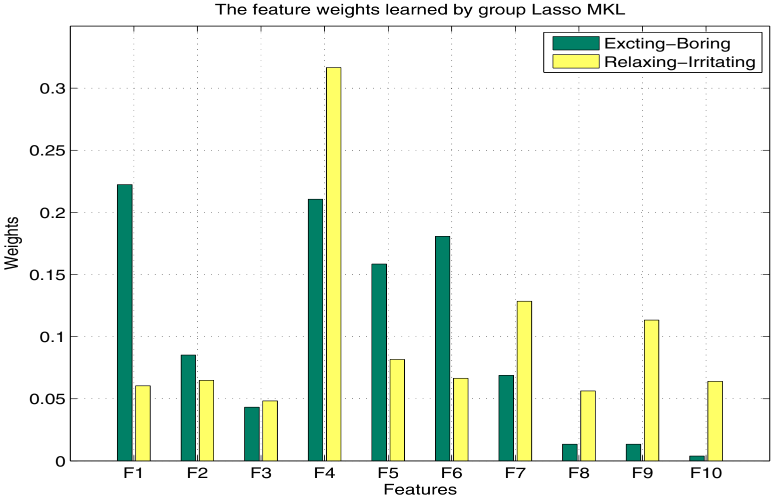


Fig. 2. The average feature representation weights over 5-fold cross-validation by using group Lasso MKL for two adjective pairs: Exciting-Boring and Relaxing-Irritating

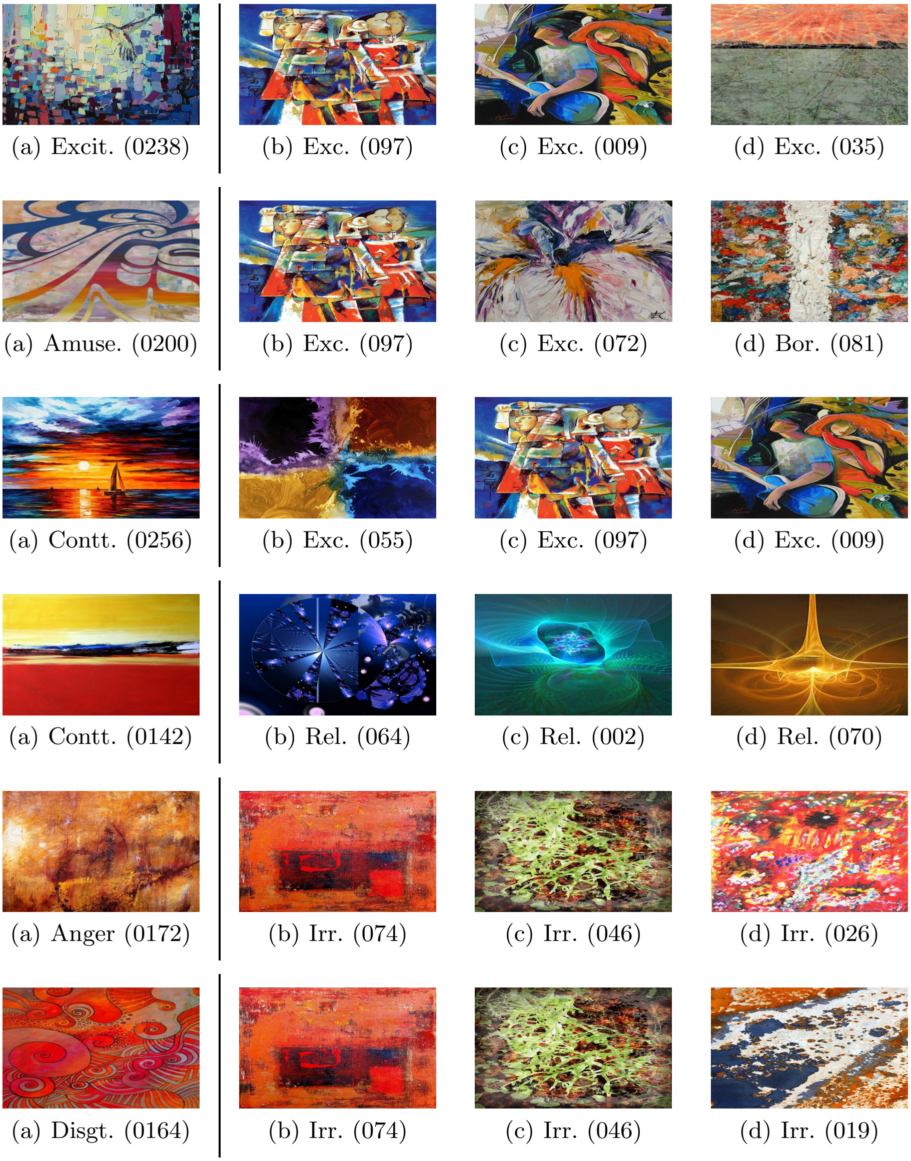
Zone5-Colm (F5) etc. For the "Relaxing vs. Irritating" pair, Zone5-Color (F4) ranks first, followed by Edge Fourier (F7), Zone5-Edgecoocc (F9), and Zone5-Colm (F5) etc. This also confirms most of the studies (e.g., [11]) that colors and edges of an image are the most informative features for affect detection. Thus, multiple kernel learning serves as a natural testbed to identify the relative importance of feature representations automatically. Table 2 shows classification results on `Abstract100` dataset. It is clear that the group Lasso MKL

Table 2. The classification performances on `Abstract100` dataset. For SVM_all/SVM_best, we conducted grid search to choose the best (C, s) pair, with $C \in (0.5,\ 1,\ 2,\ 4,\ 8)$ and $s \in (0.0078,\ 0.0156,\ 0.0312,\ 0.0625,\ 0.1250,\ 0.25,\ 0.5,\ 1,\ 2)$.

Cases/Adjective Pair	Baseline	SVM_all	SVM_best	group Lasso MKL
Exciting-Boring	0.55	0.62	0.61	**0.67**
Relaxing-Irritating	0.55	0.55	0.72	**0.73**

Table 3. The computation time (s) of the comparison methods. All the methods were implemented in MATLAB on a Macintosh computer with an Intel Core i5 processor.

Cases/Adjective Pair	Baseline	SVM (all)	SVM (best)	group Lasso MKL
Exciting-Boring	–	6.10	9.70	**0.20**
Relaxing-Irritating	–	6.01	9.70	**0.20**



(a) Excit. (0238) (b) Exc. (097) (c) Exc. (009) (d) Exc. (035)

(a) Amuse. (0200) (b) Exc. (097) (c) Exc. (072) (d) Bor. (081)

(a) Contt. (0256) (b) Exc. (055) (c) Exc. (097) (d) Exc. (009)

(a) Contt. (0142) (b) Rel. (064) (c) Rel. (002) (d) Rel. (070)

(a) Anger (0172) (b) Irr. (074) (c) Irr. (046) (d) Irr. (026)

(a) Disgt. (0164) (b) Irr. (074) (c) Irr. (046) (d) Irr. (019)

Fig. 3. The image retrieval results (displayed in “Groundtruth (index)” form) using the `Abstract280` images as queries shown in the first column, whereas the last three columns correspond to the top three retrieved images from the `Abstract100` dataset ranked by distance. The first three rows correspond to the query-retrieval results with kernel weights learned from Exciting-Boring adjective pair, whereas the last three rows correspond to the kernel weights of Relaxing-Irritating pair. Excit. = Excitement, Amuse. = Amusement, Contt. = Contentment, Disgt. = Disgust; Exc. = Exciting, Bor. = Boring, Rel. = Relaxing, Irr. = Irritating.

algorithm has achieved better classification performances than the other comparison methods in both cases. Table 3 gives the computation time of the compared methods. In either of the two cases, the computation time of group Lasso MKL is only about 1/30 of the SVM (all) and around 1/50 of the SVM (best).

4.3 Affective Abstract Image Retrieval

Experimental Setup. Both `Abstract100` and `Abstract280` datasets are used in this task. Firstly, we define the dissimilarity measure (the Euclidean distance in the implicit feature space) between a query image ($\boldsymbol{q}$) and a retrieved image ($\boldsymbol{r}$) as:

$$\begin{aligned} d_e(\boldsymbol{q},\boldsymbol{r}) &= \sqrt{k_e(\boldsymbol{q},\boldsymbol{q}) + k_e(\boldsymbol{r},\boldsymbol{r}) - 2k_e(\boldsymbol{q},\boldsymbol{r})} \\ k_e(\boldsymbol{q},\boldsymbol{q}) &= \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{q},\boldsymbol{q}) \\ k_e(\boldsymbol{r},\boldsymbol{r}) &= \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{r},\boldsymbol{r}) \\ k_e(\boldsymbol{q},\boldsymbol{r}) &= \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{q},\boldsymbol{r}) \end{aligned}$$

where $k_m(\cdot,\cdot)$ denotes the kernel function calculated on the mth feature representation and η_m is the weight for the corresponding kernel learned by the group Lasso MKL method. Therefore, given a query image $\boldsymbol{q}$, our aim is to find those images with the smallest $d_e(\boldsymbol{q},\boldsymbol{r})$ values. In essence, the smaller $d_e(\boldsymbol{q},\boldsymbol{r})$ is, the more probable that the retrieved image $\boldsymbol{r}$ evokes similar affective feelings in people. We use the `Abstract280` images as query images and let the MKL algorithm find the most relevant images from the `Abstract100` dataset. The kernel weights are selected on the complete set of `Abstract100` images (without splitting), either based on the "Exciting vs. Boring" or the "Relaxing vs. Irritating" adjective pair.

Results. Figure 3 shows the image retrieval results of certain query images for both cases. For the first case "Exciting vs. Boring", the "Excitement" image (0238) from `Abstract280` dataset successfully finds the other three "Exciting" images from `Abstract100` dataset as the first three returns. Similar results (except the "Boring" image (081)) can be observed for the "Amusement" image (0200) and the "Contentment" image (0256), due to the fact that the three emotional categories conceptually correlate with each other in the affective space [9]. For the second case "Relaxing vs. Irritating", the "Contentment" image (0142) also finds the other three "Relaxing" images as its top matches, which shows that an "Exciting" image often makes people feel "Relaxing" as well and vice versa. Both the "Anger" image (0172) and the "Disgust" image (0164) have retrieved "Irritating" images as their most relevant candidates. According to the Oxford Dictionary, the adjective word "Irritating" is defined as causing (someone) annoyance, impatience, anger, or irritation to a body part.

5 Conclusions and Future Work

In this paper, we have applied multiple kernel learning framework for affective classification and retrieval of abstract art images. MKL can make use of different feature representations or views of an image simultaneouly such that it jointly learns the feature representation weights and the corresponding classifier, which seeks for maximizing the classification performance without explicit feature selection steps. The group Lasso MKL algorithm has been adopted in the framework in that it is simple and efficient. The experimental results on two abstract image datasets have demonstrated the advantages of the group Lasso MKL in terms of feature selection, classification performance, and interpretation, for the affective abstract image classification and retrieval task.

It is worth emphasizing that MKL framework is not confined to detecting affect on abstract art images, but can be easily extended to other artistic (photographic) images and other affective stimuli such as audio and video data, given that the features and labels are available. Due to the varying subjectivity in humans and the limit of the available affective databases, it is of course not guaranteed that the MKL algorithm can make a perfect classification or retrieval for every single image. Methods such as zero-shot learning [13] may help to relieve the subjectivity and annotation issues. Eventually, the development in this interdisciplinary area relies on the joint efforts from, for instance, artificial intelligence, computer vision, cognitive science, psychology, and art theory.

Acknowledgements. This work has received funding from the Academy of Finland in the project Finnish Center of Excellence in Computational Inference Research (COIN). We gratefully acknowledge Ms Na Li for collecting the original `Abstract100` dataset.

References

1. Bach, F.: Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research 9, 1179–1225 (2008)
2. Bach, F., Lanckriet, G., Jordan, M.: Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning (ICML). ACM (2004)
3. Chang, C., Lin, C.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3) (2011)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
6. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40(2), 5 (2008)
7. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2211–2268 (2011)

8. Hanjalic, A.: Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Processing Magazine 23(2), 90–100 (2006)
9. Honkela, T., Lindh-Knuutila, T., Lagus, K.: Measuring adjective spaces. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010, Part I. LNCS, vol. 6352, pp. 351–355. Springer, Heidelberg (2010)
10. Laaksonen, J., Koskela, M., Oja, E.: PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. IEEE Transactions on Neural Networks 13(4), 841–853 (2002)
11. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the International Conference on Multimedia, pp. 83–92. ACM (2010)
12. Ou, L., Luo, M.R., Woodcock, A., Wright, A.: A study of colour emotion and colour preference. Part I: Colour emotions for single colours. Color Research & Application 29(3), 232–240 (2004)
13. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Advances in Neural Information Processing Systems (NIPS), pp. 1410–1418 (2009)
14. Shamir, L., Macura, T., Orlov, N., Eckley, D.M., Goldberg, I.G.: Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. ACM Transactions on Applied Perception 7(2) (2010)
15. Sjöberg, M., Muurinen, H., Laaksonen, J., Koskela, M.: PicSOM experiments in TRECVID 2006. In: Proceedings of the TRECVID 2006 Workshop (2006)
16. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proceedings of the 12th International Conference on Computer Vision (ICCV), pp. 606–613. IEEE (2009)
17. Wang, W., He, Q.: A survey on emotional semantic image retrieval. In: Proceedings of 15th IEEE International Conference on Image Processing, pp. 117–120 (2008)
18. Xu, Z., Jin, R., Yang, H., King, I., Lyu, M.: Simple and efficient multiple kernel learning by group lasso. In: Proceedings of the 27th International Conference on Machine Learning (ICML), pp. 1175–1182 (2010)
19. Zhang, H., Augilius, E., Honkela, T., Laaksonen, J., Gamper, H., Alene, H.: Analyzing emotional semantics of abstract art using low-level image features. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 413–423. Springer, Heidelberg (2011)

Improving Accuracy for Image Parsing Using Spatial Context and Mutual Information

Thi Ly Vu, Sun-Wook Choi, and Chong Ho Lee

School of Information and Communication Engineering,
Inha University, Incheon 402-751, Korea
vuthily@inha.edu, swchoi@inhaian.net, chlee@inha.ac.kr

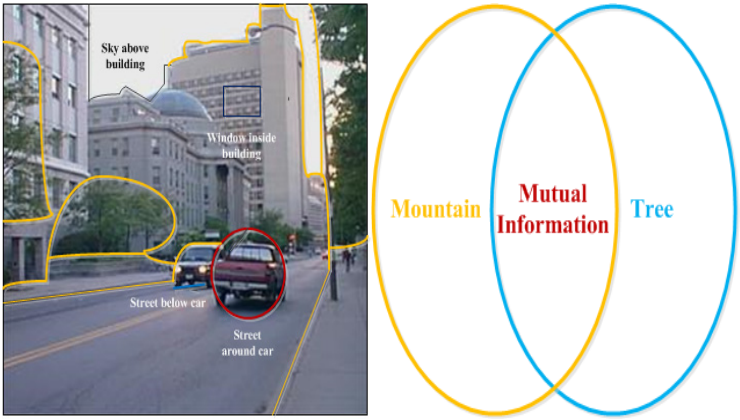
Abstract. This paper presents a novel approach for image parsing based on nonparametric model in superpixel level. Spatial context and mutual information between object co-occurrence are introduced and applied for improving the accuracy of image parsing. These methods make the probability of object co-occurrence more reliable, and thus the inference of object label from K nearest neighbors is more accurate. Our system integrates the probability of object co-occurrence with the spatial context and mutual information into a Markov Random Field(MRF) framework. Experimental results on SIFTFlow and Barcelona dataset shows that the spatial context and the mutual information are promising methods to improve the accuracy of nonparametric image parsing models.

Keywords: image parsing, MRF, superpixel, spatial context, mutual information, SIFTFlow.

1 Introduction

Image parsing is one of hard problems in computer vision. The challenging recognition task is focused in many recent works [1, 2, 4, 5, 6, 12, 14]. Firstly as noted in [5], this issue is an incredible confusing of visual words, that means one region can be matched with another region from hundreds of different labels. Secondly, the scenes are random, the objects are assumed to appear randomly leading to huge object distribution. Thirdly, due to the limit of the number of object labels in a parsing model, thus it is hard to build a completely plausible model, since the number of objects in the real scenes are actually unlimited. Recently, several works based on nonparametric models for image parsing are presented in literature [4, 5, 6, 12, 14]. Our system is inspired from the method introduced by Tighe and Lazebnik for image parsing [4] using scalable nonparametric model in superpixel levels. Their system shows pioneer result using the K nearest neighbor method which is then become the basis for other researchers for improvement. There are several directions to improve the performance of MRF model [4]. Among them, the probability of object co-occurrence are commonly utilized [5, 6]. For example, if one object label is already categorized as "street", then it has a high probability to believe that the surrounding object labels are likely "car", "sidewalk", "building" etc. Hence, taking into account this mutual

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 176–183, 2013.



(a) Spatial context (b) Mutual information

Fig. 1. Spatial context and mutual information are projected in our method

relationship, the accuracy of the image parsing model can be enhanced. The previous approaches only consider the object co-occurrence information, and do not consider the spatial relationships of objects [4, 5, 6, 12, 14]. Therefore, in the proposed approach, the spatial context is utilized to prevent this weakness. In addition, the mutual information (e.g frequency of co-occurrence objects) is meaningful for the co-occurrence probability of joint object labels as illustrated in Fig. 1 for configuring class labels in the spatial context and mutual information. Our key contribution is that the spatial context and mutual information are used to measure relationship of co-occurrence objects. Therefore, the probability of object co-occurrence becomes more reliable leading to a more accurate image parsing model. Our system achieves better accuracy rate up to 1.05% comparing to the previous methods by several experiments presented in Sect. 4.

The outline of this method is organized as below: Section 1 introduces in general image parsing reviewing previous methods and briefly describes our proposed approach. The spatial context and mutual information are detailed in Sect. 2. Section 3 presents our proposed approach of image parsing model using these two features. The experimental results are shown in Sect. 4. Finally, Sect. 5 discusses about this problem.

2 Proposed Method

2.1 Spatial Context in the Scene

In the area of object recognition and parsing image, the spatial context is also already applied for computing the weight function in previous works [1, 8, 9, 10]. In the work of C.Galleguillos *et al.* [1] the spatial context is modeled as 3D spatial descriptor to categorize object labels. In model [10], Singhal *et al.*

pre-defines the relationship between regions of image and the spatial context is presented by binary value of specific relationship for categorizing. In contrast, our system employs conditional probability between object labels having spatial context in pixel level.

As mentioned above, the spatial context is useful for capturing spatial relationship between two co-occurrence objects. To keep the scalable property of our model, the probability of location relationship is calculated by accumulating from the training set. Figure 1(a) shows a skeleton of our proposed model based on the spatial context. In our system, four location relations are defined: above, below, inside, and around. For each relation, the relation probability is estimated as:

$$P\left(l_i, l_j/r\right) \cong \frac{N\left(i, j/r\right)}{N\left(i, j\right)} \tag{1}$$

where r is the location relationship; $N\left(i, j\right)$ is the number of times pixel i and pixel j are neighbors; $N\left(i, j/r\right)$ is the number of times pixel i and pixel j are neighbors in the relation r. Then the smoothing energy of spatial context can be modeled as follows:

$$E_{spatial} = -\frac{1}{2}\log\left(\sum_{r=1}^{4} P\left(l_i, l_j/r\right)\right) \tag{2}$$

Incorporating the spatial context in our image parsing system, the performance is improved up to 0.75%. This result proved that the spatial context is promising feature to improve the performance in our model.

2.2 Mutual Information between Object Co-occurrence

Several previous works [11, 17, 18] have estimated the weight function of object labels based on mutual information. However, usage of only the mutual information to measure the probability of co-occurrence is seemly not enough because the mutual information gives out only the object co-occurrence in one image while the neighboring objects probability is not considered. Thus, in our system both information are considered to estimate the weight function. Figure 1(b) shows an illustration about the mutual information between "mountain" and "tree", that represents co-occurrence frequency of these objects in one image.

Similar to [11], the mutual information is used as the weight function, as more informative co-occurrence is provided, the feature becomes stronger. Following the method proposed in [4], the mutual information is calculated as:

$$I\left(l_i, l_j\right) = \log\frac{P\left(l_i, l_j\right)}{P\left(l_i\right)P\left(l_j\right)} = \log\frac{N_{total}freq\left(l_i, l_j\right)}{freq\left(l_i\right)freq\left(l_j\right)} \tag{3}$$

where $P\left(l_i, l_j\right)$ is the probability of co-occurrence objects l_i and l_j; N_{total} is the number of class labels; $freq\left(x\right)$ is the frequency of object x in the dataset;

$freq\left(x, y\right)$ is the co-occurrence frequency of object x and object y in the dataset. Using the mutual information, the co-occurrence energy is defined by:

$$E_{mutual} = -\log\left(P\left(l_i/l_j\right) + P\left(l_j/l_i\right) + I\left(l_i, l_j\right)\right) \tag{4}$$

where $P\left(l_i/l_j\right)$ is the conditional probability of object l_i having l_j and $I\left(l_i, l_j\right)$ is the mutual information between two classes.

In our system, the mutual information improves the performance up 0.3%. The effectiveness of combining the spatial context and mutual information is proved by several experiments where the accuracy rate is increased up to 1.05% comparing to the previous works.

3 Proposed Image Parsing Method Based on Spatial Context and Mutual Information

3.1 Retrieval Set and Superpixel Level

In image parsing problem, as mentioned in Sect. 1, the distribution of object is not uniform, hence we prefers to use nonparametric model that infers a superpixel from the most similar superpixels in the so-called retrieval set of image which contains K images most similar to the test image. Based on [14] our system also uses four types of global image features: Spatial pyramid, GIST, Tiny image and Color histogram to calculate the distance from each image in the training set to the test image. Then K images corresponding to smallest distances are selected to put into the retrieval set. An informative retrieval set should contain scene images similar to the test image.

As several approaches in image parsing area [4, 5, 6, 12, 14], the labels are assigned in the superpixel level. As mentioned in [4], this reduces computational load for this system. The superpixel is a region produced by a segmentation algorithm. In this work, the fast graph-based segmentation algorithm [3] is applied for segmenting image into superpixels; and each superpixel is presented by 20 features as in [14]. These features are calculated for every superpixel in each image to measure the distance between superpixels in the test image and superpixels in the retrieval set.

3.2 Contextual Inference

In order to enforce contextual constraints on image parsing problem, MRF model is preferred to the CRF model because the CRF model is very costly in learning and inference. Therefore, to assign label $l = \{l_1, l_2, ..., l_j\}$ to the set of superpixels $S = \{s_1, s_2, ..., s_i\}$ the per-class likelihood score of superpixels and probability of every co-occurrence object in retrieval set are put into the fully connected MRF model[11]. Similar to [4, 5, 6, 12, 14], the image labeling is formulated as minimization of standard MRF energy function defined based on labels l:

$$J\left(l\right) = \sum_{j=1:n;i=1:m} \mu\left(l_j, s_i\right) + E_{smooth} \tag{5}$$

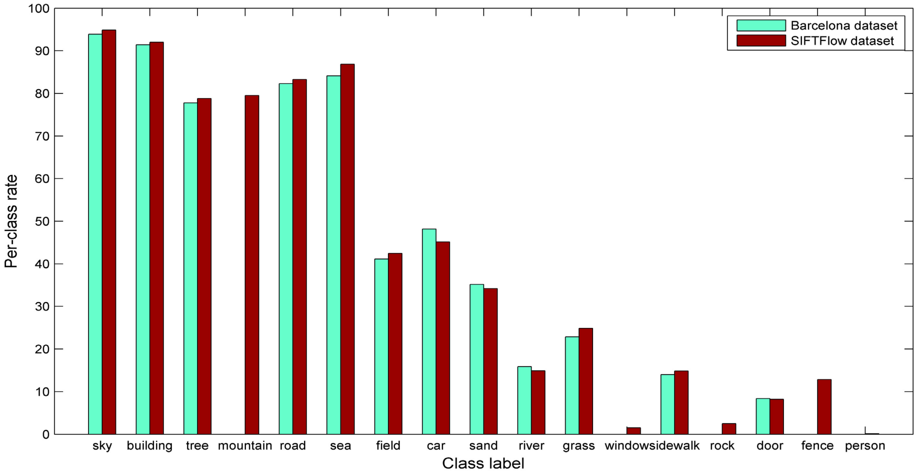


Fig. 2. Comparison per-class rate between Barcelona and SIFTFlow dataset. The Barcelona dataset has 170 labels, however we only show some labels which are common in SIFTFlow dataset to compare.

where n and m are the number of object labels in the retrieval set and superpixels in the test image; $\mu(l_j, s_i)$ presents negative logarithm of per-class likelihood scores for each superpixel s_i; smoothing term E_{smooth} shows the negative logarithm of probability between two object labels in retrieval set. This probability is calculated by accumulating the number of object co-occurrence in the dataset.

In our system, in order to increase plausibility of inference, E_{smooth} is defined by (6):

$$E_{smooth} = E_{spatial} + E_{mutual} \tag{6}$$

where $E_{spatial}$ and E_{mutual} energy from (2) and (4). These two values contain the information about probability of object co-occurrence labels including the spatial context and mutual information.

4 Experiments

For evaluating our proposed approach, several experiments on the Barcelona(a part of LabelMe dataset) and SIFTFlow datasets[19] are conducted. The Barcelonna dataset contains 14871 training images and 279 testing images in 170 labels. The SIFTFlow dataset includes 2488 train images and 200 test images in 33 labels. Figure 2 displays per-class accuracy rate for both datasets showing that our system can be executed in scalable datasets with large changes due primarily to differences in the label frequency (e.g., there are no fences in the Barcelona dataset). It is clear that classes are very non-uniform, there are few classes like "sky", "building", "tree" are very common, but some classes are rare like "people", "door". Therefore the nonparametric model is preferred for these unbalanced data sets.

Table 1. Compare accuracy rate of image parsing system on SIFTFlow dataset

Result(%)	Liu[14] 2011	J.Tighe[4] 2009	David[5] 2012	J.Young[6] 2012	J.Tighe[12] 2013	Proposed method 2013
Per-pixel	76.67	76.90	77.10	77.14	77.00	**78.19**
Per-class	x	29.38	32.5	32.29	30.10	**30.11**

As the SIFTFlow dataset is commonly used for the experiments in the previous works [4, 5, 6, 12, 14], this paper intends to compare the performance of the proposed method based on this dataset. Our result is displayed in Table 1, that shows the effectiveness of using the spatial context and mutual information on image parsing problem, the evaluation method is same as [4]. As shown in Table 1, our system achieves an overall per-pixel accuracy rate of 78.19% while the state-of-the-art per-pixel rate is 77.14% as reported in [6]. This table also shows that our per-class rate is similar to another method. Therefore, our model is a scalable image parsing method [12] but the performance is better than [12].

Example of the classified test image from our experiments are shown in Fig. 3(a, b, c, d). As seen in the figure, Fig. 3(a, b, d) show accurate results of object labeling when the spatial context and mutual information of co-occurrence object are used. To see the advantage of our system, Fig. 3(d) and Fig. 3(e) show the results of image parsing with/without using the spatial context and mutual information. In Fig. 3(e), ”building” is completely surrounded by ”field”, ”mountain”, ”plant”, ”tree” and ”sky”. This error occurs because ”building” sometimes co-occurs with ”tree”, ”plant”, ”mountain” in training images. In our model, the spatial context and frequency of co-occurrence objects are both considered. Therefore the probability of ”building” in this image becomes smaller because the ”building” completely surrounded by ”plant”, ”tree”, ”mountain” is avoided.

Clearly, the spatial context and mutual information that capture the relationship of object location and frequency of co-occurrence objects in an image is a strong visual cue. That will provide more avenues for improving categorization accuracy. Figure 3(c) shows the error case with object labels have low occurrence frequency in training dataset(”sand”).

5 Discussion

This paper has presented a novel approach for image parsing inspired by [4]. The proposed approach is not time-consuming for training except for basic computation of some statistics such as label co-occurrence probability. The accuracy of image parsing is improved in our proposed method by incorporating the spatial context and mutual information into MRF framework.

Our key contribution is that the co-occurrence relationship between object l_i and l_j can be viewed as the spatial context $P(l_i, l_j/r)$ and weighted by the

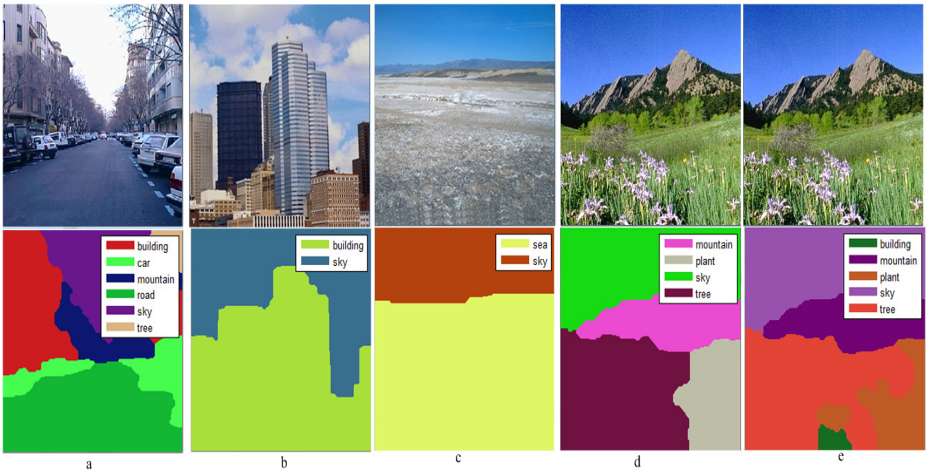


Fig. 3. Results of our parsing image system on SIFTFlow dataset. The first row is the test images, the second row is the result images. The columns (a, b, c, d) are results of our model in which column (c) shows error case when the object label has low frequency appearance in training image. The columns (d, e) compare image result between BaseMRF(e) model and our proposed approach(d). The "building" appears in Base MRF model(e) while it is not included in the test image. In our approach(d), the result does not contain the "building".

mutual information $I\left(l_i, l_j\right)$. To project all the object labels into spatial matrix, the probability of co-occurrence in each special location relation is considered. The distribution for co-occurrence objects is calculated for each relation r (e.g. above, below, inside, around) and then $E_{spatial}$ energy is computed as the sum of the distributions. The mutual information is retrieved by accumulating the frequency of co-occurrence in the dataset that gives out the distributions of l_i and l_j.

The experiment shows that our system based on simple but effective features to get the spatial context and mutual information, therefore it can be apply for various datasets. The main limitation of our system is that it uses nonparametric model hence the result sometimes depends on the retrieval set. In the future works, we will prevent this limitation by improving weight function between co-occurrence objects.

References

1. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using Co-Occurrence, Location and Appearance. In: CVPR (2008)
2. Choi, M.J., Lim, J.J., Torralba, A., Willsky, A.S.: Exploiting Hierarchical context on a large database of object categories. In: CVPR (2010)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-Based Image Segmentation. IJCV (2003)

4. Tighe, J., Lazebnik, S.: SuperParsing: Scalable Nonparametric Image parsing with super-pixels. In: CVPR (2010)
5. Eigen, D., Fergus, R.: Nonparametric image parsing using adaptive neighbor sets. In: CVPR (2012)
6. Myeong, H., Chang, J.Y., Lee, K.M.: Learning object relationships via graph-based context model. In: CVPR (2012)
7. Chow, C.K., Liu, C.N.: Approximating discrete probability distributions with dependence tree. In: IEEE (1968)
8. He, X., Zemel, R.S., Carreira-Perpinan, M.A.: Multiscale conditional random fields for image labeling. In: CVPR (2004)
9. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multiclass object recognition and segmentation by jointly modeling appearance, shape and context. IJCV (2007)
10. Singhal, A., Luo, J., Zhu, W.: Probabilistic spatial context models for scene content understanding. In: CVPR (2003)
11. Church, K.W., Hanks, P.: Words association norms, mutual information and lexicography. In: Proceedings of the 27th Annual Conference of the Association for Computational Linguistics (1989)
12. Tighe, J., Lazebnik, S.: SuperParsing: Scalable Nonparametric Image parsing with super-pixels. IJCV (2013)
13. Kumar, S., Hebert, M.: A hierarchical field framework for unified context-based classification. In: ICCV (2005)
14. Liu, C., Lim, J., Torralba, A.: Nonparametric scene parsing: Label transfer via dense scene alignment. In: IEEE (2011)
15. Tighe, J., Lazebnik, S.: Understanding scenes on many levels. In: CVPR (2011)
16. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools (2003)
17. Russell, B.C., Torralba, A., Fergus, R., Freeman, W.T.: Object regconition by scene alignment. In: NIPS (2007)
18. Weeds, J., Weir, D.: Co-occurrence retrieval: A flexible framework for lexical sistributional similarity. In: Computational Linguistic (2006)
19. Dataset for image parsing experiments, `http://www.cs.unc.edu`

Text-Prompted Multistep Speaker Verification Using Gibbs-Distribution-Based Extended Bayesian Inference for Reducing Verification Errors

Shuichi Kurogi, Takuya Ueki, Yuta Mizobe, and Takeshi Nishida

Kyushu Institute of Technology, Tobata, Kitakyushu, Fukuoka 804-8550, Japan
{kuro@,ueki@kurolab2.,mimizobe@kurolab2.}cntl.kyutech.ac.jp
http://kurolab2.cntl.kyutech.ac.jp/

Abstract. This paper presents a method of text-prompted multistep speaker verification for reducing verification errors. The method is developed for our speech processing system which utilizes competitive associative nets (CAN2s) for learning piecewise linear approximation of nonlinear speech signal to extract feature vectors of pole distribution from piecewise linear coefficients reflecting nonlinear and time-varying vocal tract of the speaker. This paper focuses on reducing verification errors by means of multistep verification using Gibbs-distribution-based extended Bayesian inference (GEBI) in text-prompted speaker verification. The effectiveness of GEBI and the comparison to BI (Bayesian inference) is shown and analyzed by means of experiments using real speech signals.

Keywords: Text-prompted multistep speaker verification, Gibbs-distribution-based extended Bayesian inference, Competitive associative net.

1 Introduction

This paper presents a method of text-prompted multistep speaker verification. Here, from [1], text-prompted speaker verification has been developed to combat spoofing from impostors and digit strings are often used to lower the complexity of processing. However, since it would be simple for today's devices to record a person saying 10 digits and to produce digits by simply typing on a keypad, text-prompted modality itself is not enough for anti-spoofing. So, we would combine other information, such as a knowledge database, other biometrics, and so on, and a specific method would be designed depending on the application. From another point of view, the present method focuses on reducing verification errors by means of multistep verification using Gibbs-distribution-based Bayesian inference (GEBI). Here, GEBI has been introduced for overcoming a problem of multistep Bayesian inference (BI) for rejecting unregistered speaker in speaker identification [5]. Note that the error rate is considered to be reduced with the increase of the number of test digits. However, this property has not been examined in detail so far, although there are experimental results showing this property, e.g. [2] shows five-digit sequences gives lower error rate than four-digit sequences.

On the other hand, the present method employs competitive associative nets (CAN2s). Here, the CAN2 is an artificial neural net for learning efficient piecewise linear approximation of nonlinear function [3]. Recently, we have shown that feature vectors of pole

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 184–192, 2013.

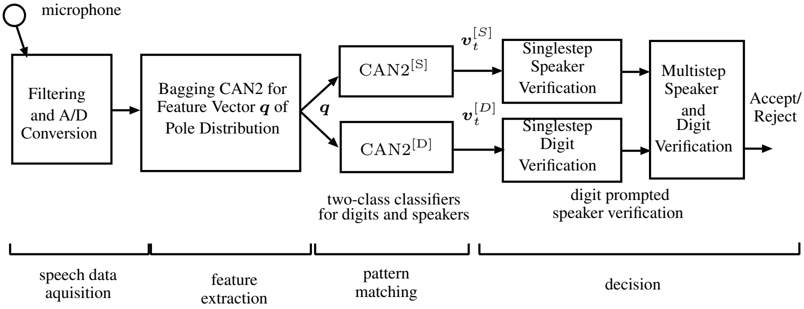

microphone
Filtering and A/D Conversion
Bagging CAN2 for Feature Vector q of Pole Distribution
q
CAN2[S]
CAN2[D]
$v_t^{[S]}$
$v_t^{[D]}$
Singlestep Speaker Verification
Singlestep Digit Verification
Multistep Speaker and Digit Verification
Accept/ Reject
two-class classifiers for digits and speakers
digit prompted speaker verification
speech data aquisition
feature extraction
pattern matching
decision

Fig. 1. Diagram of text-prompted speaker verification system using CAN2s

distribution extracted from piecewise linear predictive coefficients obtained by the bagging (bootstrap aggregating) version of the CAN2 reflect nonlinear and time-varying vocal tract of the speaker [4]. Here note that the most common way to characterize the speech signal in the literature is short-time spectral analysis, such as Linear Prediction Coding (LPC) and Mel-Frequency Cepstrum Coefficients (MFCC), where spectral features of the speech are extracted from each of consecutive interval frames spanning 10-30ms [6]. Thus, a single feature vector of LPC and MFCC corresponds to the average of multiple piecewise linear predictive coefficients of the bagging CAN2. Namely, the bagging CAN2 learns more precise information on the speech signal.

In the next section, we show an overview and formulation of singlestep speaker and digit verification system using CAN2s. And then we introduce multistep speaker and digit verification to execute text-prompted speaker verification. In order to use the same thresholds for all speakers and digits, we show a method to tune the thresholds for binarizing continuous output of CAN2s. In **3**, we show experimental results and examine the effectiveness of the present method.

2 Text-Prompted Multistep Speaker Verification System

2.1 Singlestep Digit and Speaker Verification

Fig. 1 shows an overview of our text-prompted speaker verification system using CAN2s. In the same way as general speaker recognition systems [6], it consists of four steps: speech data acquisition, feature extraction, pattern matching, and making a decision. In this research study, we use a feature vector of pole distribution obtained from a speech signal (see [4] for details). In order to achieve text-prompted speaker verification using digits, let $S = \{s_i | i \in I_S\}$ and $D = \{d_i | i \in I_D\}$ denote a set of speakers $s \in S$ and digits $d \in D$, respectively, where $I_S = \{1, 2, \cdots, |S|\}$ and $I_D = \{1, 2, \cdots, |D|\}$. Furthermore, let $\mathrm{RLM}^{[M]}$ for $M = S$ and M be a set of regression learning machines $\mathrm{RLM}^{[m]}$ $(m \in I_M)$, and each $\mathrm{RLM}^{[m]}$ learns to approximate the following target function:

$$y^{[m]} = f^{[m]}(\boldsymbol{q}) = \begin{cases} 1 \text{ if } \boldsymbol{q} \in Q^{[m]} \\ -1 \text{ otherwise} \end{cases} \tag{1}$$

In the following, we apply the same procedures for both speaker and digit processing, and we use the variables m and M for representing s and S for speakers and d and D for digits. After the learning, singlestep verification or two class classification by binarizing the output $\hat{y}^{[m]} = \hat{f}^{[m]}(\boldsymbol{q}^{[m]})$ of RLM$^{[m]}$ as

$$v^{[m]} = \begin{cases} 1 \text{ if } \hat{y}^{[m]} \geq y_\theta^{[m]} \\ -1 \text{ otherwise} \end{cases} . \tag{2}$$

Namely, we accept the speaker m if $v^{[m]} = 1$, and reject otherwise. Here, note that the threshold y_θ is tuned for stable verification as shown below.

2.2 Multistep Digit and Speaker Verification

In order to execute text-prompted speaker verification, or to execute speaker verification only when digit sequence is accepted, we employ multistep verification of digit and speaker parallelly, and then combine the result. For each multistep verification, we suppose the probability of the output vector $\boldsymbol{v}^{[M]} = (v^{[m_1]}, \cdots, v^{[|M|]})$ of binarized output of RLM$^{[M]}$ obtained for a feature vector $\boldsymbol{q} \in Q^{[m]}$ of the signal source $m \in M$ holds the conditional independence as follows:

$$p(\boldsymbol{v}^{[M]}|m) = \prod_{m_i \in M} p(v^{[m_i]}|m) \tag{3}$$

Furthermore, let $\boldsymbol{v}_{1:T}^{[M]} = \boldsymbol{v}_1^{[M]}\boldsymbol{v}_2^{[M]} \cdots \boldsymbol{v}_T^{[M]}$ be the sequences of $\boldsymbol{v}^{[M]}$ obtained for an input (speaker or digit) sequence $m_{1:T} = m_1 m_2 \cdots m_T$. In order to verify whether $\boldsymbol{v}_{1:T}^{[M]}$ is a response of a reference sequence $m_{1:T}^{[r]}$, we use GEBI (see **Appendix A.1**) and calculate two recursive posterior probabilities for $t = 1, 2, \cdots, T$ as follows,

$$p_{\mathrm{G}}\left(m_{1:t}^{[r]} \mid \boldsymbol{v}_{1:t}^{[M]}\right) = \frac{1}{Z_t} p_{\mathrm{G}}\left(m_{1:t-1}^{[r]} \mid \boldsymbol{v}_{1:t-1}^{[M]}\right)^{\beta_t/\beta_{t-1}} p\left(\boldsymbol{v}_t^{[M]} \mid m_t^{[r]}\right)^{\beta_t} , \tag{4}$$

$$p_{\mathrm{G}}\left(\overline{m_{1:t}^{[r]}} \mid \boldsymbol{v}_{1:t}^{[M]}\right) = \frac{1}{Z_t} p_{\mathrm{G}}\left(\overline{m_{1:t-1}^{[r]}} \mid \boldsymbol{v}_{1:t-1}^{[M]}\right)^{\beta_t/\beta_{t-1}} p\left(\boldsymbol{v}_t^{[M]} \mid \overline{m_t^{[r]}}\right)^{\beta_t} . \tag{5}$$

Here, Z_t is the normalization constant for holding $p_{\mathrm{G}}\left(m^{[r]} \mid \boldsymbol{v}_{1:t}^{[M]}\right) + p_{\mathrm{G}}\left(\overline{m^{[r]}} \mid \boldsymbol{v}_{1:t}^{[M]}\right) = 1$, and we employ $p\left(\boldsymbol{v}_t^{[M]} \mid \overline{m^{[r]}}\right) = \sum_{m \in M \setminus \{m^{[r]}\}} p\left(\boldsymbol{v}_t^{[M]} \mid m\right)/(|M| - 1)$.

At $t = T$, we provide the decision of T-step digit verification by

$$V_{1:T}^{[D]} = \begin{cases} 1 \text{ if } p_{\mathrm{G}}\left(d_{1:T}^{[r]} \mid \boldsymbol{v}_{1:T}^{[D]}\right) \geq p_\theta^{[D]} \\ -1 \text{ otherwise} \end{cases} \tag{6}$$

and T-step speaker and digit verification, or text-prompted speaker verification, by

$$V_{1:T}^{[\mathrm{SD}]} = \begin{cases} 1 \text{ if } \left(V_{1:T}^{[D]} = 1\right) \wedge \left(p_{\mathrm{G}}\left(s_{1:T}^{[r]} \mid \boldsymbol{v}_{1:T}^{[S]}\right) \geq p_\theta^{[S]}\right) \\ -1 \text{ otherwise} \end{cases} \tag{7}$$

where $p_\theta^{[D]}$ and $p_\theta^{[S]}$ are thresholds. Of course, $V_{1:T}^{[D]} = 1$ and $V_{1:T}^{[\mathrm{SD}]} = 1$ indicate the acceptance, and -1 the rejection.

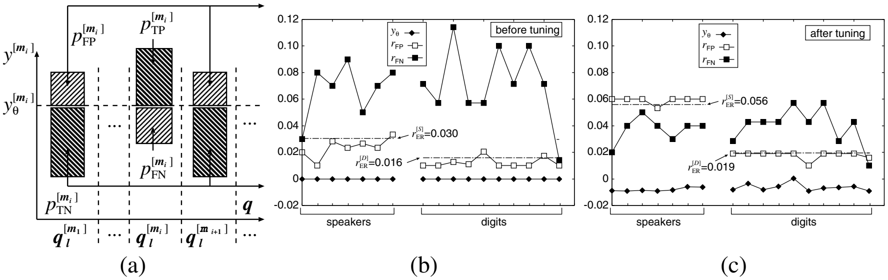


Fig. 2. (a) Schematic relationship between the threshold $y_\theta^{[m_i]}$ and the performance ratios $r_{\mathrm{FP}}^{[m_i]}$, $r_{\mathrm{FN}}^{[m_i]}$, $r_{\mathrm{TP}}^{[m_i]}$, $r_{\mathrm{TP}}^{[m_i]}$. The horizontal axis indicates all training feature vectors $\boldsymbol{q} = \boldsymbol{q}_l^{[m_i]}$ ($l = 1, 2, \cdots$) obtained from speakers $m_i = s_i \in S$ or digits $m_i = d_i \in D$. The vertical axis indicates the corresponding output $y^{[m_i]}$ of RLM$^{[m_i]}$. Experimental results (see **3**) of y_θ, r_{FP}, r_{FN} for the original $y_\theta = 0$ and tuned y_θ are shown in (b) and (c), respectively.

2.3 Tuning the Threshold y_θ of Regression Learning Machines

In order to use the same thresholds for all speakers and digits, we tune the threshold $y_\theta = y_\theta^{[m]}$ in (2) for speakers $m = s \in S$ and digits $m = d \in D$, respectively, by the following procedure. First, note that the performance of RLM$^{[m_i]}$ ($m_i \in S$) to classify the feature vectors $\boldsymbol{q}$ of m is characterized by FP (false positive) and FN (false negative). The ratio of them for randomly selected data is estimated as

$$r_{\mathrm{FP}}^{[m_i]} = \frac{1}{|M|-1} \sum_{m \in M \backslash m_i} p\left(v^{[m_i]} = 1 \mid m\right), \quad r_{\mathrm{FN}}^{[m_i]} = p\left(v^{[m_i]} = -1 \mid m_i\right), \quad (8)$$

while the ratio of TP (true positive) and TN (true negative) is $r_{\mathrm{TP}}^{[m_i]} = 1 - r_{\mathrm{FN}}^{[m_i]}$ and $r_{\mathrm{TN}}^{[m_i]} = 1 - r_{\mathrm{FP}}^{[m_i]}$. Here, note that each RLM$^{[s_i]}$ is trained to minimize the mean square error $\left\langle (v^{[m_i]} - y^{[m_i]})^2 \right\rangle$ which indicates the minimization of the error rate:

$$r_{\mathrm{ER}}^{[s_i]} = \frac{1}{|S|} \left(r_{\mathrm{FN}}^{[m_i]} + (|M|-1) r_{\mathrm{FP}}^{[m_i]}) \right). \quad (9)$$

Fig. 2(a) shows that $r_{\mathrm{FP}}^{[m_i]}$ decreases and $r_{\mathrm{FN}}^{[m_i]}$ increases for the increase of $y_\theta^{[m_i]}$. Considering this relationship, we obtain the set $\mathcal{Y}^{[m_i]}$ of triplets $\left(y_{\theta,n}^{[m_i]}, r_{\mathrm{FP},n}^{[m_i]}, r_{\mathrm{FN},n}^{[m_i]} \right)$ for trial thresholds $y_{\theta,n}^{[m_i]} = (n/n_y)\left(\overline{y}_P^{[m_i]} - \overline{y}_N^{[m_i]} \right) + \overline{y}_N^{[m_i]}$ for $n \in I_y = \{0, 1, 2, \cdots, n_y\}$, where $\overline{y}_P^{[m_i]}$ and $\overline{y}_N^{[m_i]}$ are the mean of positive and negative output $y^{[m_i]}$ for all training feature vectors $\boldsymbol{q}$, and n_y is the number of partitions. Next, we divide the data in $\mathcal{Y}^{[m]}$ ($m \in M$) into

$$\mathcal{Y}_l^{[m]} = \left\{ (y_\theta, r_{\mathrm{FP}}, r_{\mathrm{FN}}) \in \mathcal{Y}^{[m]} \;\middle|\; l = \underset{l' \in I_y}{\operatorname{argmin}} |r_{\mathrm{FP}} - r_{l'}| \right\}, \quad (10)$$

where $r_l = (l/n_y)(r_1 - r_0) + r_0$ for $l \in I_y$, $r_0 = \min\left\{ r_{\mathrm{FP},n}^{[m]} | m \in M, n \in I_y \right\}$ and $r_1 = \max\left\{ r_{\mathrm{FP},n}^{[m]} | m \in M, n \in I_y \right\}$. Then, we calculate the mean and the variance of

$(y_\theta, r_{\mathrm{FP}}, r_{\mathrm{FN}}) \in \mathcal{Y}_l^{[m]}$, which we denote $\left(\mathrm{E}_{\mathcal{Y}_l^{[m]}}(y_\theta), \mathrm{E}_{\mathcal{Y}_l^{[m]}}(r_{\mathrm{FP}}), \mathrm{E}_{\mathcal{Y}_l^{[m]}}(r_{\mathrm{FN}})\right)$ and $\left(\mathrm{V}_{\mathcal{Y}_l^{[m]}}(y_\theta), \mathrm{V}_{\mathcal{Y}_l^{[m]}}(r_{\mathrm{FP}}), \mathrm{V}_{\mathcal{Y}_l^{[m]}}(r_{\mathrm{FN}})\right)$, for each $m \in M$ and $l \in I_y$. Then, we obtain $\widetilde{l} \in I_y$ which minimizes the sum of all variances of r_{FP} and r_{FN} as

$$\widetilde{l} = \underset{l \in I_y}{\mathrm{argmin}} \sum_{m \in M} \left(\mathrm{V}_{\mathcal{Y}_l^{[m]}}(r_{\mathrm{FP}}) + \mathrm{V}_{\mathcal{Y}_l^{[m]}}(r_{\mathrm{FN}})\right). \tag{11}$$

Thus, we use the threshold $\widetilde{y}_\theta^{[m]} = \mathrm{E}_{\mathcal{Y}_{\widetilde{l}}}(y_\theta)$ for expecting smaller variance of $r_{\mathrm{FP}}^{[m]}$ and $r_{\mathrm{FN}}^{[m]}$ for all $m \in M$.

3 Experiments

3.1 Experimental Setting

We have used the speech data sampled with 8kHz of sampling rate and 16 bits of resolution in a silent room of our laboratory. They are from seven speakers (2 female and 5 mail speakers): $S = \{\mathrm{fHS}, \mathrm{fMS}, \mathrm{mKK}, \mathrm{mKO}, \mathrm{mMT}, \mathrm{mNH}, \mathrm{mYM}\}$ for ten Japanese digits $D = \{/\mathrm{zero}/, /\mathrm{ichi}/, /\mathrm{ni}/, /\mathrm{san}/, /\mathrm{yon}/, /\mathrm{go}/, /\mathrm{roku}/, /\mathrm{nana}/, /\mathrm{hachi}/, /\mathrm{kyu}/\}$. For each speaker and each digit, ten samples are recorded on different times and dates among two months. We denote each spoken digit by $x = x_{s,d,l}$ for $s \in S$, $w \in W$ and $l \in L = \{1, 2, \cdots, 10\}$, and the given dataset by $X = (x_{s,d,l} | s \in S, d \in D, l \in L)$.

In order to evaluate the performance of the present method for untrained data, we employ the following OOB (out-of-bag) estimate which is expected to have smaller bias and variance than LOOCV (leave-one-out cross-validation) [7]. For a reference speaker s_i, we make the original training dataset $Z^{[s_i]} = ((\boldsymbol{q}(x), y^{[s_i]}(x)) | x \in X)$, where $\boldsymbol{q}(x)$ is the feature vector obtained from $x \in X$, and $y^{[s_i]}(x) = 1$ if $x \in X$ is of the speaker s_i, and -1 otherwise. Next, we execute resampling with replacement to make the bags $Z^{[s_i, \alpha|Z^{[s_i]}|^\sharp, j]}$ for $j \in J^{[\mathrm{bg}]}$, where $\alpha|Z^{[s_i]}|$ indicates the number of elements in the bag for a constant α called bagsize ratio, and $J^{[\mathrm{bg}]} = \{1, 2, \cdots, |J^{[\mathrm{bg}]}|\}$ is an index set. Here, it is expected that $me^{-\alpha}$ elements in $Z^{[s_i]}$ are not in $Z^{[s_i, \alpha|Z^{[s_i]}|^\sharp, j]}$. Thus, we execute the OOB estimate of $y^{[s_i]}(x)$ by $\hat{y}^{[s_i, \mathrm{ob}]}(x) = \left\langle \hat{y}^{[s_i, j]}(x) \right\rangle_{j \in J_x^{[s_i, \mathrm{ob}]}}$, where $\hat{y}^{[s_i, j]}(x)$ is the output of $\mathrm{RLM}^{[s_i, j]}$ which has learned $Z^{[s_i, \alpha|Z^{[s_i]}|^\sharp, j]}$, and $J_x^{[s_i, \mathrm{ob}]} \triangleq \{j | (\boldsymbol{q}(x), y^{[s_i]}(x)) \notin Z^{[s_i, \alpha|Z^{[s_i]}|^\sharp, j]}, j \in J^{\mathrm{bg}}\}$. Here $\langle \cdot \rangle$ indicates the mean and the subscript indicates the range of the mean. Note that the experiments shown below are done for $|J^{[\mathrm{bg}]}| = 300$, $\alpha = 1.6$, $|Z^{[s_i]}| = |S||D||L| = 700$. For regression learning machines we use CAN2s with the number of units $N = 40$ for learning 38-dimensional feature vector $\boldsymbol{q}$ (see [4] for details of $\boldsymbol{q}$). The OOB estimate for digit verification is done by the same procedure as above.

To examine the present method, we show experimental results for a number of datasets, whether each dataset consists of 1000 pair of T-length digit sequences of test and reference speakers. Precisely, for a test digit sequence $d_{1:T} = d_1 d_2 \cdots d_T$ of a test speaker s and the corresponding reference digit sequence $d_{1:T}^{[r]} = d_1^{[r]} d_2^{[r]} \cdots d_T^{[r]}$ of a

Table 1. Experimental result of acceptance rates $r^{[D]}_{\mathrm{acc}}$ and $r^{[\mathrm{SD}]}_{\mathrm{acc}}$ achieved by the methods using (a) GEBI and tuned $\widetilde{y}_\theta$, (b) BI and tuned $\widetilde{y}_\theta$ and (c) GEBI and original $y_\theta = 0$, and four datasets with $(r_{\mathrm{CS}}, r_{\mathrm{CD}}) = (0, 1)$, $(1, 1)$, $(1, 4/5)$ and $(1, 3/5)$. The values of $r^{[D]}_{\mathrm{acc}}$ and $r^{[\mathrm{SD}]}_{\mathrm{acc}}$ are expressed by the rate [%] to the total 1000 test sequences for each case. The thresholds are $p^{[D]}_\theta = 0.96$ for (a) and (b), 0.942 for (c), $p^{[S]}_\theta = 0.5$ and $T = 15$.

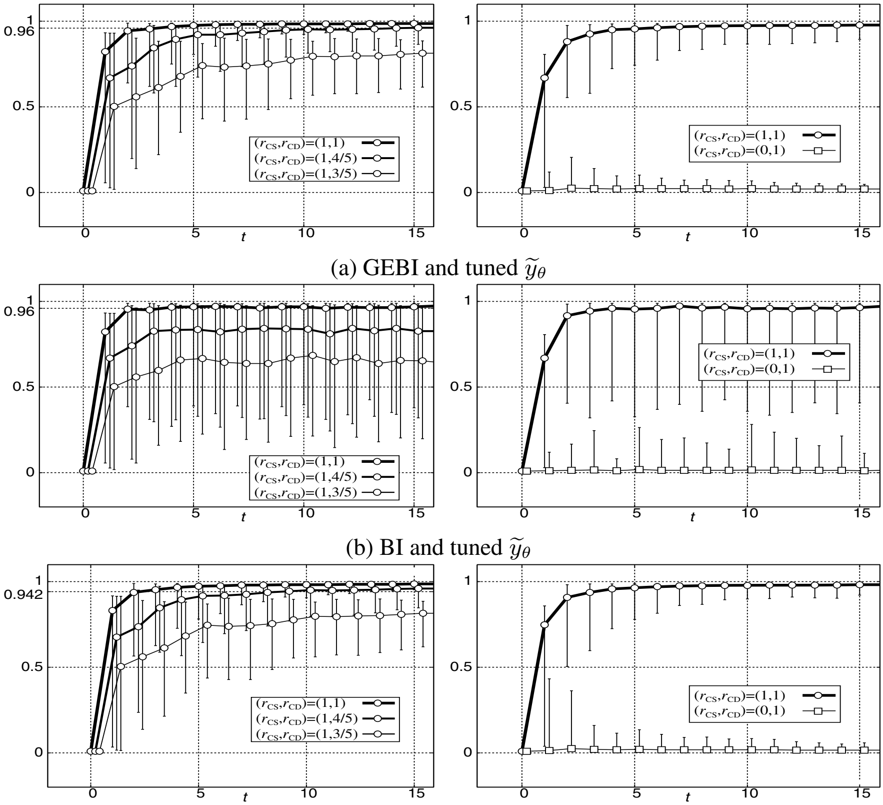
		(a) GEBI & $\widetilde{y}_\theta$		(b) BI & $\widetilde{y}_\theta$		(c) GEBI & $y_\theta = 0$	
r_{CS}	r_{CD}	$r^{[D]}_{\mathrm{acc}}$	$r^{[\mathrm{SD}]}_{\mathrm{acc}}$	$r^{[D]}_{\mathrm{acc}}$	$r^{[\mathrm{SD}]}_{\mathrm{acc}}$	$r^{[D]}_{\mathrm{acc}}$	$r^{[\mathrm{SD}]}_{\mathrm{acc}}$
0	1	**97.5**	*0.0*	**94.9**	*0.0*	**98.4**	*0.0*
1	1	**98.1**	**98.1**	**96.3**	**94.4**	**98.0**	**98.0**
1	4/5	*76.3*	*76.3*	*72.5*	*70.5*	*84.8*	*84.8*
1	3/5	*0.2*	*0.2*	*44.5*	*43.7*	*0.8*	*0.8*

reference speaker $s^{[r]}$, we select d_t, s, $d^{[r]}_t$, and $s^{[r]}$ randomly under the condition that $d_{1:T}$ involves correct digits holding $d_t = d^{[r]}_t$ with a ratio of $r_{\mathrm{CD}} = n_{\mathrm{CD}}/T$, where n_{CD} represents the number of correct digits, while we consider all T digits are of the same speaker ($r_{\mathrm{CS}} = n_{\mathrm{CS}}/T = 1$) or not ($r_{\mathrm{CS}} = 0$). Here note that for a digit $d \in D$ of a speaker $s \in S$, we use $x_{s,d,l}$ with l selected randomly.

3.2 Experimental Results and Analysis

We show an experimental result in Table 1. Here, we consider a situation that each test digit sequence consists of 5-digits, such as a date consisting of month, day and the last digit of the year. In order to avoid spoofing from impostors, we suppose that the users have previously registered their secret dates and the corresponding questions to answer the dates in the enrollment phase, and then a test speaker is prompted to answer three questions by uttering the dates in the verification phase, where the questions are selected randomly. Thus, we use $T = 15 = 5 \times 3$. The top row in Table 1 for $(r_{\mathrm{CS}}, r_{\mathrm{CD}}) = (0, 1)$ shows the result of test sequences of incorrect speakers ($r_{\mathrm{CS}} = 0$) consisting of all correct digits ($r_{\mathrm{CD}} = 1$), and we can see that all of them are rejected successfully ($r^{[\mathrm{SD}]}_{\mathrm{acc}} = 0$). The rows under the top are the results for test sequences of correct speakers ($r_{\mathrm{CS}} = 1$) involving 0, 1 and 2 incorrect digits in each 5-digit sequence for $r_{\mathrm{CD}} = 1$, $4/5$ and $3/5$, respectively. Thus, $r^{[\mathrm{SD}]}_{\mathrm{acc}}$ for $(r_{\mathrm{CS}}, r_{\mathrm{CD}}) = (1, 1)$, $(1, 4/5)$ and $(1, 3/5)$ indicates TA (true-acceptance), FA (false-acceptance) and FA, respectively, and the big $r^{[\mathrm{SD}]}_{\mathrm{acc}}[\%]$ for $(r_{\mathrm{CS}}, r_{\mathrm{CD}}) = (1, 4/5)$ is not desirable. However, the very small value $r^{[\mathrm{SD}]}_{\mathrm{acc}} = 0.2$ by the method (a) for $(r_{\mathrm{CS}}, r_{\mathrm{CD}}) = (1, 3/5)$ indicates that the test sequences involving 2 incorrect digits in each 5-digit sequence is rejected. Since it is not so easy for an impostor speaker to provide 4 correct digits in each 5-digit sequence, the present method is supposed to work for avoiding spoofing.

Now, let us examine the values in Table 1 in detail by means of multistep probabilityies shown in Fig. 3. From (a) and (c), we can see that the probability curves of GEBI for $r_{\mathrm{CS}} \neq 0$ increases and the variance decreases with the increase of t. From the probability curves for digits shown on the left in (a) and (c), we can see that the



(a) GEBI and tuned $\widetilde{y}_\theta$

(b) BI and tuned $\widetilde{y}_\theta$

(c) GEBI and original $y_\theta = 0$

Fig. 3. Experimental result of multistep probability for digits (left) and speakers (right). The plus and minus error bars indicate RMS (root mean square) of positive and negative errors from the mean, respectively. The curves for different datasets are shifted slightly and horizontally to avoid crossovers.

bottom curves and the error bar ranges are below the threshold $p_\theta^{[D]} = 0.96$ and 0.942, respectively, at $t = 15$. Furthermore, from the curves for speakers shown on the right in (a) and (c), we can see that the upper curve for $r_{\rm CS} = 1$ and the lower curve for $r_{\rm CS} = 0$ can be separated by the threshold $p_\theta^{[S]} = 0.5$. Therefore, we can understand the values of $r_{\rm acc}^{\rm [SD]}$ by the methods (a) and (c) for $(r_{\rm CS}, r_{\rm CD}) = (1, 3/5)$ in Table 1 are very small. The threshold $p_\theta^{[D]} = 0.942$ for (c) in Table 1 is set smaller than 0.96 for (a) in order for $r_{\rm acc}^{\rm [SD]}$ to be almost the same value, where smaller value is necessary for (c) because the error range of the corresponding probability on the left in Fig. 3(c) is suppoed to be slightly wider owing that the tuned threshold $\tilde{y}_\theta$ used in (a) achieves the smaller variance of classification error ratios (see Fig. 2(b) and (c)).

For $(r_{\rm CS}, r_{\rm CD}) = (1, 1)$ in Table 1, the false rejection rate (FRR) is given by FRR= $100{-}r_{\rm acc}^{\rm [SD]}$ [%]. Precisely by the method (a) for the increase of $t = 5, 10, \cdots, 50$, we have obtained decreasing FRR= 10.0, 4.3, 1.9, 1.7, 1.1, 1.0, 0.5, 0.4, 0.4, 0.3, respectively.

Furthermore, for $(r_{\rm CS}, r_{\rm CD}) = (1, 3/5)$, we have obtained decreasing false acceptance rate (FAR) as FAR= $r_{\rm acc}^{[\rm SD]}$=0.5, 0.3, 0.2 and 0 for $t = 5, 10, 15$ and 20, respectively. This monotonically decreasing property is considered to be obtained from that the probability of GEBI, $p_{\rm G}\left(m|\boldsymbol{v}_{1:t}^{[M]}\right)$ given by (15) shown below, converges monotonically for the increase of t. On the other hand, the probability of BI, $p_{\rm B}(m|\boldsymbol{v}_{1:t}^{[M]})$ given by (13), also converges monotonically but slowly. Namely, $p_{\rm G}\left(m|\boldsymbol{v}_{1:t}^{[M]}\right)$ reaches convergence at t satisfying $\tilde{L}_{1:t}^{[m]} \gg |\log p_0(m)|/t$, but $p_{\rm B}\left(s|\boldsymbol{v}_{1:t}^{[M]}\right)$ does not reach convergence at t satisfying $\tilde{L}_{1:t}^{[m]} \gg |\log p_0(m)|/t$. From Fig. 3 (b), we can see that the error ranges of the curves are large and fluctuate. Therefore, we can see that $r_{\rm acc}^{[\rm SD]}$ in Table 1 for (b) using BI indicates worse performance.

4 Conclusion

We have presented a method of text-prompted multistep speaker verification using GEBI for reducing verification errors. We have shown that the probability of GEBI is more stable and reduces verification error rates much more than BI by means of the analysis of probabilities and experimental results using real speech signals. This paper considers only registered test speakers, while additional method is supposed to be necessary for unregistered test speakers, which is for our future research study.

This work was supported by JSPS KAKENHI Grant Number 24500276.

A Appendix

A.1 Multistep BI and GEBI

We briefly show a problem of multistep BI (Bayesian inference) and introduce Gibbs-distribution-based extended Bayesian inference (GEBI) (see [5] for details): For an output sequence $\boldsymbol{v}_{1:t}^{[M]} = \boldsymbol{v}_1^{[M]}\boldsymbol{v}_2^{[M]}\cdots\boldsymbol{v}_t^{[M]}$ responding to an input signal source $m \in M$ (m and M represent s and S or d and D shown in **2.1**), we can estimate the posterior by the naive BI as

$$p_{\rm B}\left(m|\boldsymbol{v}_{1:t}^{[M]}\right) = \frac{1}{Z_t} p_{\rm B}\left(m|\boldsymbol{v}_{1:t-1}^{[M]}\right)\, p\left(\boldsymbol{v}_t^{[M]}|m\right) \;. \tag{12}$$

where Z_t is the normalization constant for holding $\sum_{m\in M} p_{\rm B}(m|\boldsymbol{v}_{1:t}^{[M]}) = 1$. From this equation for $t = 1, 2, \cdots$, we have

$$p_{\rm B}\left(m|\boldsymbol{v}_{1:t}^{[M]}\right) = \frac{1}{Z_t}\exp\left(-t\left(\tilde{L}_{1:t}^{[m]} - \frac{1}{t}\log p_0(m)\right)\right) \;, \tag{13}$$

where $p_0(m) = p_{\rm B}(m|\boldsymbol{v}_{1:0}^{[M]})$ denotes the prior, and $\tilde{L}_{1:t}^{[m]} \triangleq -\frac{1}{t}\left(\sum_{k=1}^{t}\log p\left(\boldsymbol{v}_k^{[M]}|m\right)\right)$is the normalized negative log-likelihood. Then, the ratio of the probability of $m_i \in M$ to $m_\nu = \underset{m_i\in M}{\operatorname{argmax}}\, p_{\rm B}(m_i|\boldsymbol{v}_{1:t}^{[M]})$ becomes

$$r_{\mathrm{B},i,\nu} \triangleq \frac{p_{\mathrm{B}}(m_i|\boldsymbol{v}_{1:t}^{[M]})}{p_{\mathrm{B}}(m_\nu|\boldsymbol{v}_{1:t}^{[M]})} = \frac{p_{\mathrm{B}}(m_i)}{p_{\mathrm{B}}(m_\nu)} \exp\left(-t(\tilde{L}_{1:t}^{[m_i]} - \tilde{L}_{1:t}^{[m_\nu]})\right) \to \begin{cases} 1, m_i = m_\nu \\ 0, m_i \neq m_\nu \end{cases} \tag{14}$$

for $t \to \infty$, because $\sum_{m \in M} p_{\mathrm{B}}(m|\boldsymbol{v}_{1:t}^{[M]}) = 1$. Thus, $p_{\mathrm{B}}(m|\boldsymbol{v}_{1:t}^{[M]})$ becomes very large for a registered $m = m_\nu$ even when the current signal is of an unregistered source. To overcome the problem, let us use the following Gibbs distribution:

$$p_{\mathrm{G}}\left(m|\boldsymbol{v}_{1:t}^{[M]}\right) \triangleq \frac{1}{Z_t} \exp\left(-\beta\left(\tilde{L}_{1:t}^{[m]} - \frac{1}{t}\log p_0(m)\right)\right) \ , \tag{15}$$

where $p_0(m) = p_{\mathrm{G}}(m|\boldsymbol{v}_{1:0}^{[M]})$ indicates the prior, and β is a parameter called inverse temperature. Then, for the increase of t, the ratio of $p_{\mathrm{G}}(m|\boldsymbol{v}_{1:t}^{[M]})$ for $m = m_i$ to $m_\nu = \underset{m_i \in M}{\mathrm{argmax}}\, p_{\mathrm{G}}(m_i|\boldsymbol{v}_{1:t}^{[M]})$ converges to a constant value less than 1 as follows;

$$r_{\mathrm{G},i,\nu} \triangleq \frac{p_{\mathrm{G}}(m_i|\boldsymbol{v}_{1:t}^{[M]})}{p_{\mathrm{G}}(m_\nu|\boldsymbol{v}_{1:t}^{[M]})} \to \exp\left(-\beta(\tilde{L}_{1:t}^{[m_i]} - \tilde{L}_{1:t}^{[m_\nu]})\right) \to c_i^\beta < 1 \ . \tag{16}$$

Thus, we can avoid the above problem of multistep BI. Here, from (15), we derive the following stepwise inference,

$$p_{\mathrm{G}}\left(m|\boldsymbol{v}_{1:t}^{[M]}\right) = \frac{1}{Z_t} p_{\mathrm{G}}\left(m|\boldsymbol{v}_{1:t-1}^{[M]}\right)^{\beta_t/\beta_{t-1}} p\left(\boldsymbol{v}_t^{[M]}|m\right)^{\beta_t} \ , \tag{17}$$

where $\beta_t = \beta/t$ $(t \geq 1)$ and $\beta_0 = 1$. Note that the conventional BI is given by $\beta_t = 1$ $(t \geq 0)$, and we name this inference by GEBI.

References

1. Beigi, H.: Fundamentals of speaker recognition. Springer-Verlag New York Inc. (C) (2011)
2. Melin, H., Lindberg, J.: Prompting of Passwords in Speaker Verification Systems, Fonetik-97, Phonum 4, Umera University, Sweden, May 28-30 (1997)
3. Kurogi, S., Ueno, T., Sawa, M.: A batch learning method for competitive associative net and its application to function approximation. In: Proc. SCI 2004, vol. V, pp. 24–28 (2004)
4. Kurogi, S., Mineishi, S., Sato, S.: An analysis of speaker recognition using bagging CAN2 and pole distribution of speech signals. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (eds.) ICONIP 2010, Part I. LNCS, vol. 6443, pp. 363–370. Springer, Heidelberg (2010)
5. Mizobe, Y., Kurogi, S., Tsukazaki, T., Nishida, T.: Multistep speaker identification using Gibbs-distribution-based extended Bayesian inference for rejecting unregistered speaker. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012, Part V. LNCS, vol. 7667, pp. 247–255. Springer, Heidelberg (2012)
6. Campbell, J.P.: Speaker Recognition: A Tutorial. Proc. the IEEE 859, 1437–1462 (1997)
7. Kurogi, S.: Improving generalization performance via out-of-bag estimate using variable size of bags. J. Japanese Neural Network Society 16(2), 81–92 (2009)

Dynamics of Neuronal Responses in the Inferotemporal Cortex Associated with 3D Object Recognition Learning

Reona Yamaguchi[1], Kazunari Honda[1], Jun-ya Okamura[1], Shintaro Saruwatari[2],
Jin Oshima[2], and Gang Wang[1,*]

[1] Department of Information Science and Biomedical Engineering, Graduate School of Science and Engineering, Kagoshima University, Japan
{k1868980,k7756509}@kadai.jp, jokamura@ibe.kagoshima-u.ac.jp,
gwang@ibe.kagoshima-u.ac.jp

[2] Department of Information Science and Biomedical Engineering, Faculty of Engineering, Kagoshima University, Japan
{k0428951,k8763048}@kadai.jp

Abstract. Discrimination of objects at the same viewing angles develops view-invariant object recognition in some extent. To reveal the underlying neuronal mechanism, we investigated the activities of the inferotemporal cell populations responding to object images with different prior experiences. With different object sets, the monkeys were trained beforehand with the Object task in which view-invariant object recognition across similar objects was required, and the Image task in which only the discrimination at the same viewing angles was required. We found, in the level of cell population, that the responses to the images with the prior experience of the Image task were similar to those to the images with the prior experience of the Object task. The highest level in response similarity was found 260 ms after the stimulus onset. The results suggest that the view-invariant neuronal representations developed through the experience of the object discrimination at the same viewing angles.

Keywords: inferotemporal cortex, object recognition, monkey.

1 Introduction

We can recognize objects even if they are seen from different viewing angles. The capability to recognize objects across changes in the viewing angle develops as the viewer repeatedly sees both the object and distractors in rotation. It has been proposed that different views of an object become associated when they are experienced in succession during rotation [1-2]. However, our previous behavioral studies demonstrated that, in monkey, the discrimination of objects at the same viewpoint was enough for the formation of perceptual tolerance in the range of viewing angle change up to 60 deg [3]. It was shown further that the fine object discrimination experience at each of several viewing angles was required for the development of such ability for

[*] Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 193–199, 2013.

the discrimination across similar objects, whereas coarse discrimination experience didn’t generate any cross-view object discrimination ability [4]. Single-cell recordings from the macaque inferotemporal (IT) cortex have shown that cells respond to views of familiar objects and show moderately broad tunings for viewing angles [5]. Theoretical studies suggested the view-tuned units converge onto the view-invariant units [6]. It was also noted that the IT cells show broad tunings for viewing angles of 3D objects in untrained animals [7]. In the present study, we investigated the responses of cell population in the monkey inferotemporal cortex to the object images with the prior experience of discrimination of similar objects at the same viewing angles.

2 Method

We used two male macaque monkey (*Macaca fuscata*) weighting 6.5 and 7.5 kg, respectively. All procedures were performed in accordance with the guidelines of the Japan Neuroscience Society and were approved by the Animal Experiment Committee of Kagoshima University.

2.1 Visual Stimuli

Stimulus objects were created using three-dimensional graphics software (shade 9; e-frontier, Tokyo, Japan). Details of the object creation have been described previously [3]. In brief, we created four artificial objects by deforming a prototype in four different directions in three-dimensional feature space. Seven parameters of the object shape were combined into three parameters that spanned the entire feature space. In order to create the four different views, each object was rotated with a 30-deg interval around an axis perpendicular to the visual axis that connected the viewer’s eyes and the object. Each stimulus set consisted of 16 images (4 views × 4 objects).

Six object sets (set A-F) were generated from 6 prototypes which were distinct from each other. The similarity for a pair of object images across sets was significantly lower than that for any pair of objects in the same set. The size of object image was 6.5 deg on average. We used human psychophysics to make the difficulty of discrimination within each set comparable among different stimulus sets, with the percentage of correct responses ~80%.

2.2 Training Task

Before the electrophysiological recording, object images were exposed to monkeys extensively in the training session, during which the monkeys were asked to make discrimination between the objects. In the process for object discrimination, the stimulus image presentation was controlled in different ways in three tasks, i.e. the Image task, the Object task, and the Exposure task. Consistent across the three tasks, a trial started with monkey’s lever press, which turned on a fixation spot at the center of the screen. After continuous pressing the lever and fixating for 500 ms, the first stimulus appeared. Two to five stimuli were presented in each trial. Each stimulus was presented for 500 ms with 500 ms inter-stimulus intervals. The monkey had to maintain eye fixation with an accuracy of ±2.5 deg and keep pressing the lever until the last stimulus appeared. The monkey had to release the lever within 1 s when the object changed to get a juice reward. When the monkey made an incorrect response or broke the eye fixation, the trial was aborted with a beep sound. The inter-trial interval was 1.5 s after a correct response and 2.5 s after an incorrect response.

In the Exposure task, after an identical image was presented 1-4 times, an image selected from a different object set appeared. No discrimination was required between the images in the same object set. In the Image task, the images were exposed to the monkey without the opportunity to associate different views of each object. Monkeys were trained to discriminate among the images within each of the four viewing angles. After an identical image was presented 1-4 times, the image of the other object in the same set at the same viewing angle appeared. The Object task required the association across different views of each object. After different views of the same object were presented 1-4 times randomly, the image of the other object in the same set appeared.

In the training session, object set A and B, set C and D, and set E and F were used in the Exposure task, the Image task, and the Object task, respectively, for Monkey K. For Monkey H, the object sets were swapped across tasks. Object set C and F, set A and E, and set B and D were used in the Exposure task, the Image task, and the Object task, respectively.

2.3 Single Cell Recoding

After the monkey's performance in training session reached a saturation level, we conducted single cell recordings from the inferotemporal cortex. Recordings were conducted with tungsten electrodes (FHC, Bowdoinham, ME, USA), which passed through a guide tube and were advanced by a micro-manipulator (Narishige, Tokyo, Japan). The recoding sites were determined with reference to MRI images taken before the first preparatory surgery. Cells were recorded from a ventrolateral region of the inferotemporal cortex, lateral to the anterior middle temporal sulcus, in the posterior/anterior range between 18 and 26 mm anterior to the ear bar position for monkey K and between 16 and 19 mm for monkey H. All recordings were conducted while the monkey was performing the Exposure task.

2.4 Analyses of the Neural Data

We analyzed neuronal responses to the first stimulus presentation in each trial. Only those for trials with correct responses were included. The responses in the time windows of 60-560 ms, 60-240 ms and 240-420 ms from the stimulus onset were defined as the responses in the whole period, early and late phases, respectively. The magnitude of the responses was determined as mean firing rate during the whole period (60-560 ms), early phase (60-240 ms) and late phase (240-420 ms) minus the spontaneous firing rate during the 500 ms period immediately preceding the stimulus presentation. For each neuron, the significance of the response was tested with one-way ANOVA

($p < 0.05$) in the three time windows. For the population analysis, we calculated a population response vector for each image. The population response vector for the images consisted of the mean firing rates of the cells that showed statistically significant stimulus selectivity among the 16 images in the set. We measured the correlation coefficient between the population response vectors to evaluate the similarity between the population responses to each image. We compared correlation coefficients between the response to the images separated by the viewing angles of 30, 60 and 90 deg of the same object (Cs) and different objects (Cd) in a set.

3 Results

We recorded single unit activity of 353 inferotemporal cells. Among the 353 cells, 201 cells showed significant excitatory responses (84 cells for monkey K, 117 cells for monkey H) in the whole period, 177 cells showed significant excitatory responses (70 cells for monkey K, 107 cells for monkey H) in the early phase and 197 cells showed significant excitatory responses (80 cells for monkey K, 117 cells for monkey H) in the late phase.

Fig. 1 shows the averaged Cs and Cd values for the responses to the objects experienced in the Exposure task, Image task and Object task. In the whole period, at the viewing angle difference of 30 deg, two-way ANOVA with the factors of task (Exposure, Image and Object) and correlation coefficient (Cs, Cd) was performed to combined data of the two monkeys. The main effect of the factor of task (df = 2, F = 3.72, $p < 0.05$) and the main effect of the factor of the correlation coefficient (df = 1, F = 29.95, $p < 0.0001$) were significant. Also, the interaction between the factors (df = 2, F = 3.25, $p < 0.05$) were significant. Post-hoc test with Bonferroni correction showed statistically significant larger Cs value than Cd value in the responses to the objects experienced in the Image task and in those in the Object task (Image task: $p < 0.001$, Object task: $p < 0.0001$), whereas there was no significant difference between the Cs and Cd values in the responses to the objects experienced in the Exposure task. Furthermore, Cd value for the objects experienced in the Object task were significantly smaller than the Cd values for the objects experienced in the Exposure task and the Image task, respectively (Object task vs. Exposure task: $p < 0.0001$; Object task vs. Image task: $p < 0.01$). Similarly, in the late phase, at the viewing angle difference of 30 deg, the Cs value was significantly larger than the Cd value for the objects experienced in the Image task and Object task (Image task: $p < 0.0001$, Object task: $p < 0.0001$). However, in the early phase, no significant differences between the values could be confirmed at the viewing angle differences of 30, 60 and 90 deg.

The data collected from the whole period and the late phase showed similar result, with clear difference from those from the early phase. To examine the time course of the Cs and Cd for each image, the Cs and Cd values were calculated in a 40 ms time window sliding from the stimulus onset to 1000 ms after stimulus onset with a step of 20 ms. The 201 cells showing significant excitatory responses in the whole period were pooled here. The bold horizontal bars in Fig. 2 represent the time bins in which the Cs value was significantly larger than the Cd value ($p < 0.05$, unpaired *t*-test).

The difference between the Cs and Cd values started to be significant at 260 ms bin for the sets with the prior experience of Image task at the view separation of 30 deg. For the object sets with the prior experience in the Object task, the Cs became significantly larger than Cd 200 ms after the stimulus onset for both the cases of 30-deg view separation and 60-deg view separation. For the object set with the prior experience of the Exposure task, there was no significant difference between the Cs and Cd values at the view separation of 30, 60 and 90 deg.

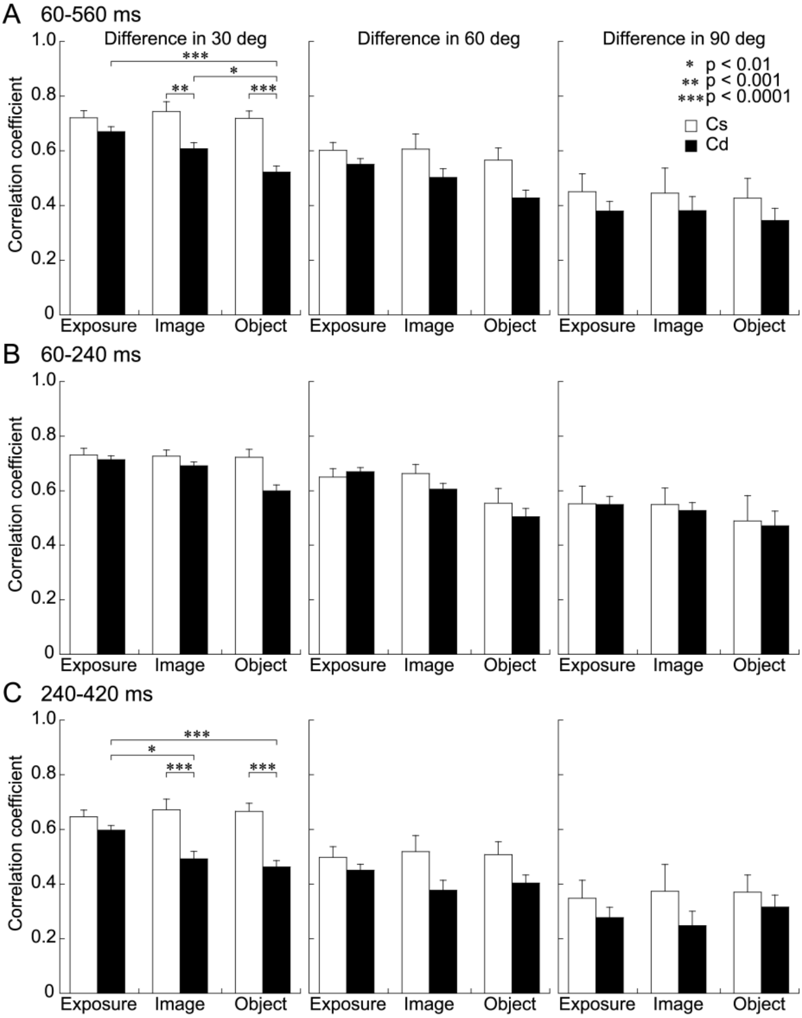


Fig. 1. The Cs and Cd values in the three time windows

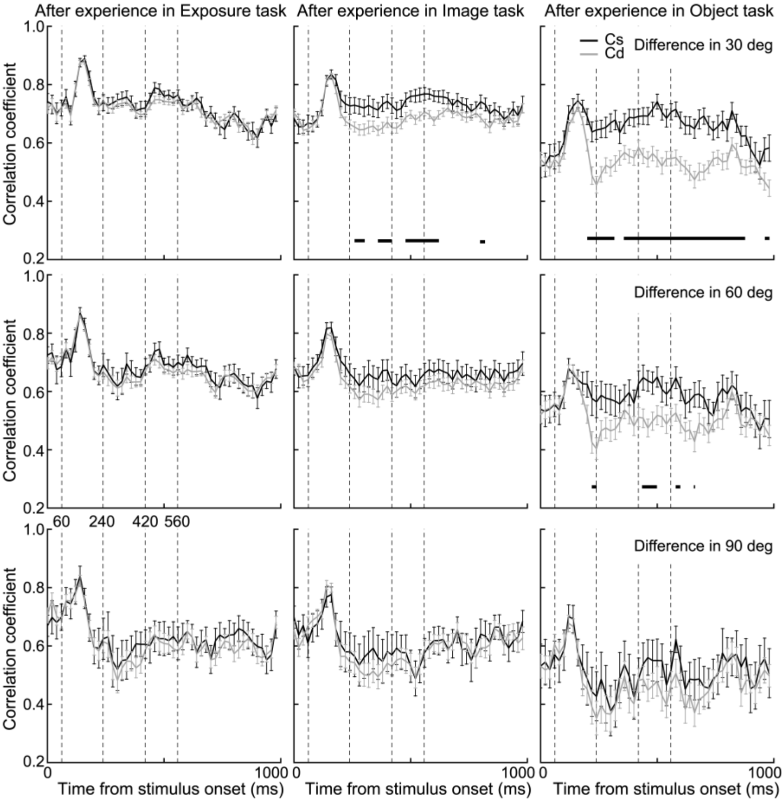


Fig. 2. Time courses of the Cs and Cd values

4 Discussion

In the present study, population responses of the IT neurons to the object images with different prior experiences were compared. The correlation coefficients between the population responses to 30-deg separated views of the same object (Cs) were significantly larger than those between the population responses to views of different objects (Cd), after the monkeys experienced the images in the Image task. Such difference between Cs and Cd was also observed in the population responses to the object images with the prior experience of the Object task, but not in the responses to the images with the prior experience of the Exposure task. The results thus suggest that the different views of the same object experienced in the Image task were represented in much more similar manner than the representations for the views of different objects. In the Image task, discrimination between similar objects, even if it was at the same viewpoints, generated the representations of the experienced views for the same objects, with differentiation of the representations for the experienced different objects. The generation for views of the same objects and differentiation for views of different objects were dependent on the extent of requirement for the prior discrimination experience, since the results were limited to the fine discrimination (the Image task) but not the coarse discrimination (the Exposure task). The experience dependent change in the representation of object image was similar to that generated by the Object task which required the association of the views for the same object. The differences between the Cs and Cd were not significant at the viewing angle differences in 60 deg and 90 deg in the population responses to the object images experienced in the Image task, Object task and Exposure task. Further analysis may be necessary in order to examine the baseline of the correlation coefficients and the preference of the neurons to the object images experienced, since the sensitivity of the IT neurons changes depending on the discrimination experience of the objects [8].

Temporally, the Cs value didn't differ significantly from Cd value in the early phase of the responses no matter of the prior experience on the object images. At the viewing angle difference of 30 deg (Fig. 2), the difference between the Cs and Cd values became significant at 260 ms and 200 ms after the stimulus onset for the responses to the object images with the prior experiences of the Image task and the Object task, respectively. It remained significant in almost all the late phase for both of the cases, but in the case with the prior experience of the Object task the significance between the Cs and Cd values kept until about 900 ms after stimulus onset. Considering the time necessary for the visual information to reach the inferotempral cortex, it is reasonable to think that the response in the early phase should mainly represent the bottom-up information from the early cortical areas. Therefore, the formation of view-invariant object recognition should not be completed in the earlier cortical stages. The convincing significance between the Cs and Cd values in the late phase suggest the similar representations for views of the same objects and distinct representations for different objects in the time period of 240-420 ms, which should contribute to the formation of view-invariant object recognition ability.

References

1. Földiák, P.: Learning invariance from transformation sequences. Neural Comp. 3, 194–200 (1991)
2. Stryker, M.P.: Temporal associations. Nature 354, 108–109 (1991)
3. Wang, G., Obama, S., Yamashita, W., Sugihara, T., Tanaka, K.: Prior experience of rotation is not required for recognizing objects seen from different angles. Nat. Neurosci. 8, 1768–1775 (2005)
4. Yamashita, W., Wang, G., Tanaka, K.: View-invariant object recognition ability develops after discrimination, not mere exposure, at several viewing angles. Eur. J. Neurosci. 31, 327–335 (2010)
5. Logothetis, N.K., Pauls, J., Poggio, T.: Shape representation in the inferior temporal cortex of monkeys. Curr. Biol. 5, 552–563 (1995)
6. Riesenhuber, M., Poggio, T.: Models of object recognition. Nat. Neurosci. 3, 1199–1204 (2000)
7. Hung, C.C., Carlson, E.T., Connor, C.E.: Medial axis shape coding in macaque inferotemporal cortex. Neuron 74, 1099–1113 (2012)
8. Kobatake, E., Wang, G., Tanaka, K.: Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. J. Neurophysiol. 80, 324–330 (1998)

Multiple Metric Learning for Graph Based Human Pose Estimation

Mohammadreza Zolfaghari[1], Morteza Ghareh Gozlou[2],
and Mohammad Taghi Manzuri Shalmani[1]

[1] Sharif University of Technology,
Department of Computer Engineering, Tehran, Iran
[2] University of Surrey,
Faculty of Engineering and Physical Sciences, Guildford, United Kingdom
mzolfaghari@ce.sharif.edu, m.gharehgozlou@sitechltd.com,
manzuri@sharif.edu

Abstract. In this paper, a multiple metric learning scheme for human pose estimation from a single image is proposed. Here, we focused on a big challenge of this problem which is; different 3D poses might correspond to similar inputs. To address this ambiguity, some Euclidean distance based approaches use prior knowledge or pose model that can work properly, provided that the model parameters are being estimated accurately. In the proposed method, the manifold of data is divided into several clusters and then, we learn a new metric for each partition by utilizing not only input features, but also their corresponding poses. The manifold clustering allows the decomposition of multiple manifolds into a set of manifolds that are less complex. Furthermore, the input data could be mapped to a new space where the ambiguity problem is minimized. Our guiding principle for learning the distance metrics is to preserve the manifold structure of the input data. The proposed method employs Tikhonov regularization technique to obtain a smooth estimation of the labels. Experiments on the data set of human pose estimation demonstrate that the proposed multiple metric learning consistently outperforms single-metric learning method across different activities by a wide margin.

Keywords: Multiple metric learning, semi-supervised estimation, human pose estimation.

1 Introduction

3D human pose estimation from monocular images has been a well-studied topic in the computer vision. This problem faced with some challenges such as many different poses may have similar image descriptors. Human pose estimation becomes even more challenging if the image descriptors cannot be properly detected due to self-occlusion or presence of complex background. Effective solutions for these difficulties will affect performance of many applications such as

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 200–208, 2013.

video surveillance, activity recognition, motion capture, and human-computer interaction. In general, there are three schools of thought in this area as follows:

The model-based methods employ a parametric body model, based on prior knowledge and estimate the configuration of human body by optimization methods. These approaches require both good initialization and proper models. In addition, the model-based methods have got high computational cost to find the solution. Moreover, these methods may trap into sub-optimal solutions [1].

The learning-based methods utilize a direct mapping between the input and output spaces to confront the demand for initialization, precise body modeling, and other difficulties [2,3]. These methods are attractive because many learning techniques exist for pose estimation in real-time applications. One drawback of the learning-based methods is that their performance depends on the size of the training data.

The example-based methods store a set of training data that corresponding poses are known, and then these methods use a similarity measure to find the most similar training data to the unknown test input [4,5]. The main problem of these methods is their need to perform a query, both quickly and reliably. In addition, we should incorporate enough examples to reach a good performance.

Both the learning-based and example-based approaches, such as nearest neighbors, regression and mixture of experts, face a big challenge: different 3D poses might correspond to similar inputs. Some methods use the available knowledge regarding the output space to deal with this situation, while there are difficulties such as learning lots of parameters and incorporating large training sets to cover the variability in people appearances [6]. Euclidean distance based approaches cannot solve this problem because they are influenced by the distance metric.

Recently, several attempts have been made to reduce this problem by presenting new distance metric. We can make Euclidean distance more useful by learning a linear transformation of variables with the goal that for each example, examples of the same classes stay near together and examples with different classes becoming far from each other. One popular solution is Large Margin Nearest Neighbors (LMNN) [7], which learns the distance metric while the k-nearest neighbors belong to the same class and data points from dissimilar classes separated by a large margin. The LMNN approach and the most of other metric learning approaches are specially designed for classification problems. However, human pose estimation is a regression problem and the constraints in the LMNN method and previous research are not feasible for separating the data points of different classes [8]. To the best of our knowledge, [5] and [3] are only metric learning approaches for human pose estimation. However, [5] proposed an example-based method to human pose estimation. Hence, it needs lots of training data to cover the variations and [3] learns only one metric for whole input spaces without considering manifold structure and input space complexity.

In this paper, we present a multiple metric learning approach to overcome aforementioned problems by partitioning the data manifold into a set of manifolds, and then we learn distance metrics for each manifold by minimizing quadratic objective function. We use the label information of the data not only in manifold construction, but also in multiple metric learning. Since the labels provide important cues about similarities among samples, learning metrics with this approach decreases ambiguity and provides better distance metric, at which similar samples tend to stay close and dissimilar data points become far from each other, particularly under a small given dataset.

The rest of the paper is organized as follows: In Section 2, we present our multiple metric learning approach. Section 3, demonstrates the experimental results and finally, Section 4 provides the conclusions and outlines future work.

2 Proposed Method

In this paper, we use a graph based semi-supervised approach for 3D human pose estimation. We assume that the data lies on a low dimensional manifold, and the labels change smoothly on this manifold. Manifold assumption is held in many real world applications such as human pose estimation due to the fact that input features of human shapes captured from human activities, have a small degree of freedom [3]. Thus, the labels change smoothly over the manifold with little changes in the feature space. The manifold structure is estimated under the assumption that the whole data space is locally Euclidean, hence Euclidean distance is used to understand the local structure of the original space. Consequently, we cannot estimate the manifold structure properly and the data manifold roughly bends over itself, particularly when the data dimension is high and the number of data is not enough. Therefore, we cannot estimate the manifold structure properly and the data manifold bends close to itself, particularly when the data dimension is high and the number of data is not enough. This situation contradicts with the assumption that the labels variations are smooth on the manifold and causes reduction of the performance. Using good judgment, Euclidean distance based approaches cannot handle this problem because they are influenced by the distance metric.

Empirical studies [10] show that learning multiple metrics from the data can improve the performance of methods. In this paper, a multiple metric learning has been considered, where the data manifold is partitioned into several manifolds, and then we learn a distance metric for each manifold independently. Our aim is to map the data into a space, where the projected data does not have the previously mentioned problems, so the semi-supervised pose estimation could obtain better accuracy. Experiments show that even we can obtain better results than MTIK [3] with a simple manifold partitioning method such as k-means. Experiments show that we can even obtain better results than MTIK [3] with a simple manifold partitioning method such as k-means. Fig. 1(a) visualizes the data manifold (using a 3-nearest neighbor graph) related to part of "Walk" activity before the mapping. The manifold partitions are colored with different colors, where the edges connect similar points. We have reduced the dimensionality of the original space to 3 for presentation purposes. Fig. 1(b) visualizes the estimated manifold of the same data in a similar manner after mapping through the proposed multiple metric learning. As it can be seen, the data manifold in the new space approximately bends over itself in fewer places.

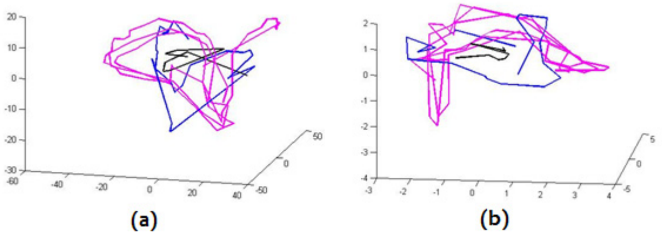


Fig. 1. The estimated manifold of the data with three clusters. Figures (a) and (b) show the manifold before and after mapping respectively.

2.1 Multiple Metric Learning

In this section, we, firstly, focus on the single metric learning problem. Then we propose a multiple metric learning method and the optimization function. The learning method takes a training set of N observations, $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where $x_i \in R^d$ are the input features and $y_i \in R^k$ are their labels in the pose space. The goal of the metric learning problem is to find a transformation matrix A, after applying which, the distance between two inputs x_i and x_j may be measured as:

$$d_A(x_i, x_j) = \sqrt{(x_i - x_j)^T A (x_i - x_j)}, \tag{1}$$

where A is a symmetric positive definite matrix ($A \succeq 0$), is known as the Mahalanobis distance matrix. The eq. (1) gives the Euclidean metric if $A = I$, the identity matrix. The distance $d_A(x_i, x_j)$ should preserves the local neighborhood property that two data points stay close if their labels are similar. For this purpose, we introduce the regularization function $H(A)$ as follows:

$$H(A) = \sum_{i,j} \frac{(\zeta_{ij} + 1)}{2} (d_A^2(x_i, x_j) - \hat{d}^2(i, j))^2, \tag{2}$$

where $\zeta_{ij} = 1$ if y_i and y_j are connected in a k-nearest neighbor graph constructed in the pose space, and $\zeta_{ij} = -1$ otherwise. Also, $\hat{d}(i, j)$ is the measure of the desired distance that [3] and [8] defined as follows:

$$\hat{d}(i, j) = (\frac{\alpha ||y_i - y_j||_2 + \gamma}{C - ||y_i - y_j||_2 + \epsilon})^p \times ||x_i - x_j||_2. \tag{3}$$

In the desired distance measure, we utilize the label information to modify the distance between data by a factor of the variation tendency of data (i.e. if the distance between the labels of two data is large, then the desired distance is also large, vice versa). Here, α, γ, which denotes the labeling noise, and p, making data easier to discriminate, are three constants, $C = \max_{i,j}\{||y_i - y_j||_2\}$, and $\epsilon >$

0 which ensures the denominator never equals zero. The single metric learning problem is to minimize the following cost:

$$\begin{aligned} \min_{A} \quad & H(A), \\ s.t. \quad & \frac{1}{m}\sum_{i,j}\frac{(\zeta_{ij}-1)}{2}d_A^2(x_i - x_j) \le \tau, \\ & A \succeq 0, \end{aligned} \tag{4}$$

Where $\tau > 0$ is the margin and m is the number of dissimilar points. The second term in (4) puts the dissimilar data points by a margin in the new space. In other words, the distance of two similar data in the new space should be small by a factor of $\hat{d}(i,j)$ and larger than τ otherwise.

Suppose we cluster the original manifold of data into K manifolds using classical k-means algorithm. For each cluster $k \in K$, we are given m_k examples $\{(x_1^k, y_1^k), \ldots, (x_{m_k}^k, y_{m_k}^k)\}$. We have now reduced the original problem into K smaller problems, for each we learn a distance metric. Our multiple metric learning algorithm achieves this goal by minimizing the following objective:

$$\begin{aligned} \min_{A_1,\ldots,A_K} \quad & \frac{1}{K}\sum_{k=1}^{K} H(A_k) \\ s.t. \quad & \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{i,j}\frac{(\zeta_{ij}-1)}{2}d_{A_k}^2(x_i^k - x_j^k) \le \tau \\ & A_k \succeq 0, k = 1, \ldots, K. \end{aligned} \tag{5}$$

For each individual partition, the matrix A_k is symmetric and positive semi-definite and can be decomposed as $A_k = L_k^T L_k$. We utilize these linear transforms to project the whole data from the original input space to a new one by:

$$x^{new} = \frac{1}{\sum\limits_{k=1}^{K} w_k}(w_1 L_1^T x + w_2 L_2^T x + \ldots + w_k L_K^T x), \tag{6}$$

where $w_k = \frac{1}{d^2(x, c_k)}$ and c_k is center of the k-th partition. After mapping, we find the labeling function with respect to the mapped data. In the following we will explain the smooth labeling function over the manifold in the new space.

2.2 Regression on Manifold

Before estimating the human pose by a graph based semi-supervised method, we would require a manifold construction method. To model the manifold, we use k-nearest neighbor in the new space to connect each point to k of its nearest neighbors. Then the pose estimation can be done by finding a labeling function, where we combine Tikhonov regularization term and the error term, and solve the optimization problem [12,3]:

Table 1. Comparison of methods for different activities

Activity	Train #	RVM	TGP	TIK	MTIK	MMTIK	Imp. %
Acrobatic	100	56.37	15.49	5.98	5.74	**5.56**	3.1
	200	5.86	5.37	5.09	5.08	**5.02**	1.2
	400	5.16	4.69	4.81	4.67	**4.39**	6.0
Golf	100	22.90	6.79	5.09	4.99	**4.91**	1.6
	200	5.47	4.69	4.92	4.70	**4.63**	1.4
	400	4.24	**4.16**	4.88	4.66	4.65	0.2
Laugh	100	24.80	17.72	3.95	3.90	**3.77**	3.3
	200	4.88	4.67	3.16	3.08	**2.92**	5.1
	400	4.36	4.19	2.82	2.75	**2.71**	1.4
Walk	100	26.36	11.71	3.53	3.37	**3.26**	3.2
	200	3.68	3.58	2.96	2.91	**2.83**	3.1
	400	3.12	3.01	2.61	2.56	**2.50**	2.3

$$\min_{f \in R^z} \sum_{i \in T} (y_i - f(x_i))^T (y_i - f(x_i)) + \lambda trace(F^T L F), \tag{7}$$

Where T represents the training data, y_i is a true pose for input x_i, $f(x_i)$ is an estimated pose, λ is a positive trade-off parameter balancing the smoothness and exactness of labels, F is a matrix where $F_{i,.} = f(x_i)$, and matrix $L = D - W$ is the graph Laplacian of the manifold in the new space. Here, W is the adjacency matrix of the graph and D is a diagonal matrix, $D_{ii} = \sum_j W_{ji}$. The above optimization problem finds a trade-off between reconstruction error for each training data (first term) and smoothness of labels over the manifold (second term). We solve this problem by setting the derivative of the function (7) (with respect to f) equal to zero.

3 Experimental Result

In this section, we present experiments on the human pose data set. In all of our experiments, we used the average (over all angles) root mean square difference and automatically determined the parameters α and p with 2-fold cross validation (for speed reasons). λ, γ, and ϵ were set to values in order of 10^{-4}. We set $k = 3$ in k-NN for manifold construction and finding nearest neighbors. We cluster the manifold of input space into 3 partitions and then learn the metrics in these clusters separately. We used the histograms of shape contexts as described in [2] to encode silhouette shapes as 100-D descriptors x and 3D body model with 19 joints which results in a 57-D pose y. For each activity, we took 600 frames and used specific number of them as a training set, and used the rest as

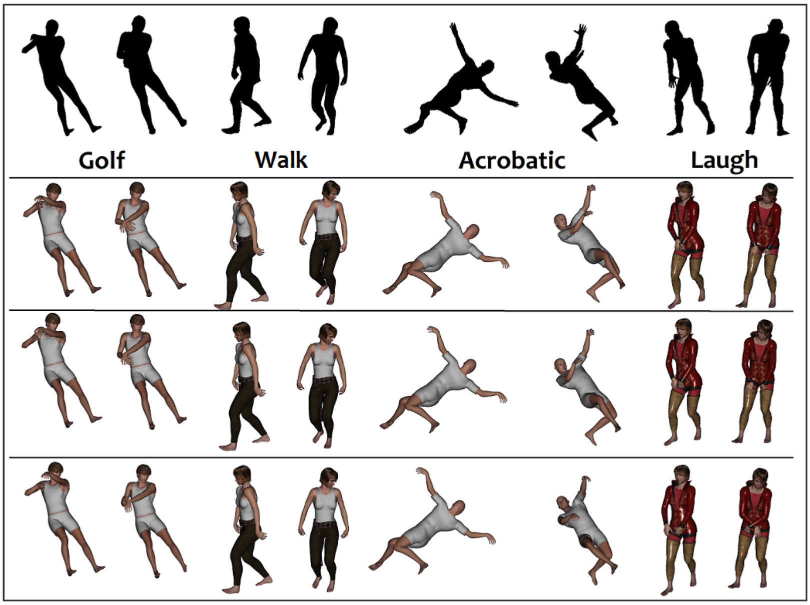


Fig. 2. Some sample 3D pose estimation for various activities not included in the training set: The rows show input *silhouettes*, *ground-truth* and outputs of our algorithm (*MMTIK*), and *MTIK*, respectively. The columns show *"Golf"*, *"Walk"*, *"Acrobatic"* and *"Laugh"* activities, respectively.

a test data set. In all of the sequences, we processed each frame independently without considering any temporal consistency. In our experiments, we compared proposed method (MMTIK) against recent proposed metric learning method for human pose estimation, namely MTIK [3], the Relevance Vector Machine (RVM) [2], the Twing Gaussian Process (TGP) [11], and Tikhonov regularization (TIK) [12]. We computed the estimation error, using four activities "Golf", "Walk", "Acrobatic", and "Laugh" of the CMU Mocap data set [9]. Numeric results are shown in Table 1. We illustrate the improvement ratio of the MMTIK with respect to MTIK in the last column of Table 1 for comparison purpose. We compute the percentage of the improvement by:

$$\frac{(e_{MTIK} - e_{MMTIK})}{e_{MTIK}} \times 100, \tag{8}$$

where e_{MTIK} and e_{MMTIK} are MTIK and MMTIK errors (in degrees) respectively. The performance of RVM and TGP is dependent on the number of training data points, thus with 100 training data points the estimation error of these algorithms is dramatically high, as it can be seen in Table 1. This table indicates how a certain amount of training data points might have influence on the performance of the methods. The presented method generally performs better than MTIK utilizing a graph-based approach to human pose estimation, in all activities with different number of training samples, with an average 2.65% in improvement ratio. A few sample images of the qualitative result using 200 training data points for various activities are shown in Fig. 2. The rows show input *silhouettes*, *ground-truth* and outputs of our algorithm (*MMTIK*), and *MTIK*, respectively. The columns show *"Golf"*, *"Walk"*, *"Acrobatic"* and *"Laugh"* activities, respectively. Notice that MMTIK has successfully reconstructed the test input, except for the hands in the "golf" and "acrobatic" sequences.

4 Conclusion

We proposed a new method for multiple metric learning which utilizes the label information to project the data to a new space where the ambiguity problem of 3D human pose estimation is reduced. We partition the manifold into several manifolds, and then learn one metric for each partition. This is an extension of existing metric learning method (MTIK) from single metric to multiple metric learning. Our experiments on different activities, show that our multiple metric learning algorithm performs significantly better than the other state-of-the-art approaches. Future work includes learning multiple metrics independently, while a shared metric between all of partitions will be learned. Finding a proper algorithm to partition the manifold of data is also part of the future research.

References

1. Lee, M., Nevatia, R.: Human pose tracking in monocular sequence using multilevel structured models. IEEE Trans. Pattern Anal. Mach. Intell. 31, 27–38 (2009)
2. Agarwal, A., Triggs, B.: Recovering 3D Human Pose from Monocular Images. IEEE Trans. Pattern Anal. Mach. Intell. 28, 44–58 (2006)
3. Pourdamghani, N., Rabiee, H.R., Zolfaghari, M.: Metric learning for graph based semi-supervised human pose estimation. In: 21th IEEE International Conference on Pattern Recognition, pp. 3386–3389. IEEE Press, Tsukuba (2012)
4. Jiang, H.: 3D Human Pose Reconstruction Using Millions of Exemplars. In: 20th IEEE International Conference on Pattern Recognition, pp. 1674–1677. IEEE Press, Istanbul (2010)
5. Jain, P., Kulis, B., Grauman, K.: Fast image search for learned metrics. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press, Anchorage (2008)
6. Agarwal, A., Triggs, B.: Monocular Human Motion Capture with a Mixture of Regressors. In: IEEE Conference on Computer Vision and Pattern Recognition, p. 72. IEEE Press, San Diego (2005)
7. Weinberger, K.Q., Saul, L.K.: Distance Metric Learning for Large Margin Nearest Neighbor Classification. J. Mach. Learn. Res. 10, 207–244 (2009)
8. Xiao, B., Yang, X., Xu, Y., Zha, H.: Learning distance metric for regression by semidefinite programming with application to human age estimation. In: 17th ACM International Conference on Multimedia, pp. 451–460. ACM, Beijing (2009)
9. Carnegie Mellon University Motion Capture Database, `http://mocap.cs.cmu.edu`

10. Cui, Z., Li, W., Xu, D., Shan, S., Chen, X.: Fusing Robust Face Region Descriptors via Multiple Metric Learning for Face Recognition in the Wild. In: 26th IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press, Oregon (2013)
11. Bo, L., Sminchisescu, C.: Twin Gaussian Processes for Structured Prediction. Int. J. Comput. Vision 87, 28–52 (2010)
12. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. J. Mach. Learn. Res. 7, 2399–2434 (2006)

Motion Deblurring Using Super-Sparsity

Jingxiong Zhao, Haohua Zhao, Keting Zhang, and Liqing Zhang*

MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems,
Department of Computer Science and Engineering,
Shanghai Jiao Tong University, China
{soyscut,haoh.zhao,zzsnail,lqzhang}@sjtu.edu.cn

Abstract. Motion blur is caused by the camera shake during the exposure in which the blur kernel describes the trace of shaking. Based on this generating process of the kernel , we observed that the distribution of the kernel obeys super-sparsity, as the natural images. Recent works mostly exploit various kinds of priors in their models, but focus on the the speed or a close-form formulation for convenience of mathematical calculation ignoring the intrinsic feature of the kernels and images. In this paper we propose a new model with super-sparse prior for the deblurring problem from one single image. Since the close-form formulation of this model doesn't exist, we use a look-up table trick to approximate the solution. Qualitative and quantitative evaluation demonstrate that our model with super-sparse prior can produce stable and high-quality results.

Keywords: Motion Deblur, Blind Deconvolution, Super-sparsity.

1 Introduction

Motion is one of the most common causes for image blurring. In many cases, camera shake during the exposure time leads to the motion blur on one image. How to recovery the sharp image from one single blurred image is a chronic fundamental problem in the area of image enhancement.

However, motion deblurring is also a very challenging problem due to the loss of information during deblurring. Typically, the blurring process can be modeled as the motion blur kernel which describes the trace of the sensor convolved with the latent image. In this way, removing the motion blur can be formulated as deconvolution problem which can further separated to blind and non-blind cases. In non-blind deconvolution, the underlying PSF is known and the task is to recover the latent image given the kernel. While blind deconvolution needs to estimate both the kernel and the latent image. The single observed blur image can result in many possible combinations of the latent image and the blur kernel if no extra constraints are given.

In the literature,the early works mainly focus on the constraints of the blur kernel[6,7,8]. Recently,some remarkable progress has been made through making

[*] To whom all correspondence should be addressed.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 209–216, 2013.

use of the natural images' statistics. Fergus et al.[1] used a mixture Gaussian model with the first-order derivative of the natural images. Shan et al.[2] built a Bayesian estimated frame for the noise with the sparse prior of natural images' multi-order derivatives. Some models[4,5,9] focus on the kernel estimation since the technics of non-blind deblurring is splendid and the accurate estimation of the PSF becomes the key for this problem. Cho et al.[4] made a rough prediction of the salient edges with the shock filter and the bilateral filter. Joshi et al.[5] favored the salient edges with the edge profile.Li et al.[9] propose a kernel estimated method using the iterative support detection kernel refinement.

Generally speaking, the success of blind deconvolution attributes to the use of various priors to avoid falling into the suboptimal solution. Some priors are studied like ℓ_1 norm([2, 11]), ℓ_2 norm[4,5] and total variation regularizer[12]. However, the prior distribution is often selected on the basis of mathematical convenience rather than as a reflection of any prior beliefs. Actually, the natural images have much heavier tails than the Laplacian(1-norm) and the Gaussian distribution as shown in Fig.1.

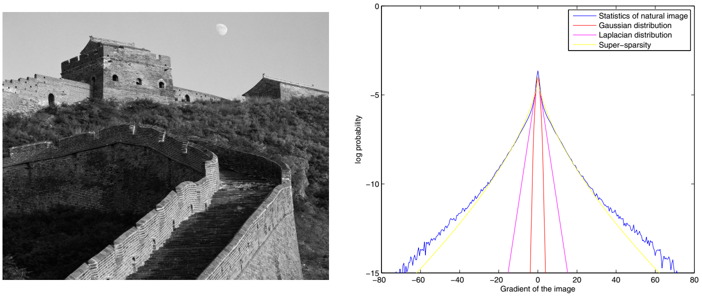


Fig. 1. A comparison between different distribution and the statistics of the natural iamge. In the right, the distributions are modeled as $y \propto e^{-k|x|^{\alpha}}$. For the Gaussian distribution, $\alpha = 2$, for the Laplacian distribution, $\alpha = 1$ and for the super-sparsity distribution, $\alpha = 1/3$ is chosen here.

Also, the kernel which blurs the latent image is usually generated by the shake of the camera during the exposure which is just a blink. Taking consideration of this process, the kernel is far more sparser than commonly used 1-norm or 2-norm prior as shown in Fig.2

In this paper, we propose a method which utilizes the super-sparse feature of natural images and the kernels based on above observations. Because there doesn't exist a close-form formulation, we'll approximate the solution through Newton-Raphson method. Considering the ranges of the variables in the formulations are limited, we use a look-up table trick to whittle down the redundant computing and accelerate the process. The corresponding result can be obtained by seeking the nearest item and linear or bilinear interpolation. Experimental results demonstrate the effectiveness and efficiency of our model.

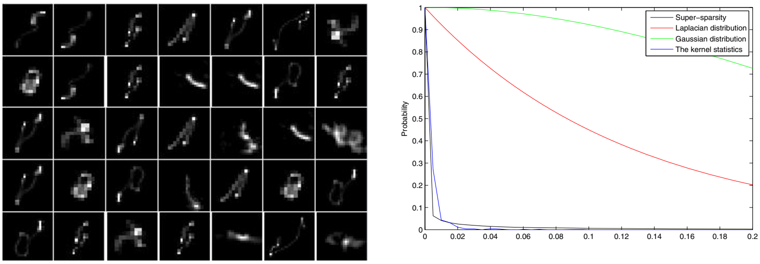


Fig. 2. A comparison between different distribution and the statistics of typical motion kernels. In the right, the distributions are modeled as $y \propto e^{-k|x|^{\alpha}}$. For the Gaussian distribution, $\alpha = 2$, for the Laplacian distribution, $\alpha = 1$ and for the super-sparsity distribution, $\alpha = 1/10$ is chosen here.

2 Model

The shift-invariant blurring problem is usually modeled as a convolution process as following:

$$B = L * K + N \tag{1}$$

where B is a blurred image, K is the blur kernel, L is a Latent image,N is the added noise and $*$ is the convolution operator. We need to estimate both kernel K and sharp image L in the situation that only B is known.

To reconstruct K and L, a simple maximum a posteriori(MAP) estimation can be used. By Baysian theorem, the process can be represented as:

$$p(L, K|B) \propto p(B|L, K)p(L)p(K) \tag{2}$$

where $p(B|L, K)$ is the likelihood term and $p(L)$ and $p(K)$ are the priors of the latent image and the kernel. To restore the image and kernel is equivalent to minimize the cost function $-logp(L, K|B)$. To obtain accurate kernel and underlying image, in the blind-deconvolution L and K can be alternatingly optimized as following:

$$L^{'} = \arg\min_{L} \|K * L - B\|^2 + \rho(L) \tag{3}$$

$$K^{'} = \arg\min_{K} \|K * L - B\|^2 + \rho(K) \tag{4}$$

In Eqs. (3) and (4), $\rho(L)$ and $\rho(K)$ are the prior term, $\|K * L - B\|^2$ is the likelihood term. As we can see, the process to solve this ill-posed problem has been divided into two phases: Kernel estimation and non-blind deconvolution. We'll describe these two phases separately.

2.1 Kernel Estimation

In the kernel estimation, the latent image L is assumed to be known. So the estimated L has effect on the kernel restoration. Here we first use a bilateral filter to smooth the image and reserve the edge, then use a shock filter to enhance the edge. Then, we can find the best kernel through minimizing the cost function:

$$C(K) = \sum_{(\nabla L, \nabla B)} \omega \|K * \nabla L - \nabla B\|^2 + \gamma \|K\|_\alpha^\alpha \tag{5}$$

where $\nabla L = \{\partial_x L, \partial_y L\}$ and $\nabla B = \{\partial_x B, \partial_y B\}$, $\omega = \{\omega_1, \omega_2\}$ represents the weights for partial derivatives in two directions. The term $\gamma \|K\|_\alpha^\alpha$ is the prior of the kernel. As mentioned above, the distribution of the kernel is far more sparser than the regularization in existed models. Based on the statistics of the motion blur, $\alpha = 0.1$ will be used as the super-sparse prior term in the experiments.

To solve Eq. (5), Bregman Iteration is put into use. The auxiliary variable Φ should be introduced, then the new cost function can be divided into two part and minimized respectively, one for Φ and one for K:

$$C(\Phi) = \gamma \|\Phi\|_\alpha^\alpha + \lambda \|\Phi - K\|_2^2 \tag{6}$$

$$C(K) = \sum_{(\nabla L, \nabla B)} \omega \|K * \nabla L - \nabla B\|^2 + \lambda \|\Phi - K\|_2^2 \tag{7}$$

With known Φ, the optimal solution for K can be solved by $\partial C(K)/\partial K = 0$. To solve the Eq. (7) with convolution efficiently, Fourier transform can be introduced according to the Plancherel's theorem.The solution of K can be expressed as:

$$K = \mathcal{F}^{-1}\left(\frac{\omega \mathcal{F}(\nabla B) \circ \overline{\mathcal{F}(\nabla L)} + \lambda \mathcal{F}(\Phi)}{\omega \mathcal{F}(\nabla L) \circ \overline{\mathcal{F}(\nabla L)} + \lambda}\right) \tag{8}$$

where $\overline{(\cdot)}$ is the complex conjugate and $\circ$ represents the component-wise operator, the division is also element-wise.

With known K, the Eq. (6) with super sparse term is obviously a non-convex problem. So It only has approximate solution. Here we use a Newton-Raphson method to solve the equation numerically. For the convenience of data reused and to accelerate the process of solving the problem at every pixel, one look-up table's trick will be deployed.

In the Eq. (6), only K is unknown for given parameters γ and λ and fixed α. What's more, each pixel value of K has a range indeed because the sum of all pixels in K is equal to $1(0 \leq K_\tau \leq 1$, where τ is the coordinate). The numerical solution can be obtained from 0 to 1 in a fixed interval(0.0001 is used).These solutions will be organized to construct the lookup table. After that, for an input pixel value of K, the corresponding Φ can be solved by linear interpolation in the lookup table. Considering the large scale of reduplication for pixel values and reuse of the table, this trick can reduce the redundancy efficiently.

2.2 Non-blind Deconvolution

In this step, the kernel K is known. Similar to the kernel-estimation stage, the MAP problem can also be transformed to a cost minimization problem:

$$C(L) = \sum_{(\nabla L, \nabla B)} \left(\omega \|\nabla L * K - \nabla B\|_2^2 + \lambda \|\nabla L\|_\alpha^\alpha + \mu \|\nabla L - \nabla B\|_2^2 \circ M \right) \quad (9)$$

where $(\nabla L, \nabla B) \in \{(\partial_x L, \partial_x B), (\partial_y L, \partial_y B)\}$. M is a 2-D binary mask as a local prior term to supress the ringing artifacts which is introduced in [2]. As mentioned above, the prior of the image's gradient is actually more sparser than 1-norm and 2-norm which are used in most models. In the Eq. (9), α will also choose to be much smaller than 1(1/3 in the experiment).

Similarly, the Bregman iteration process can be deployed. To minimize the cost function, we introduce one variable Φ and divide the function into two part which will be optimized alternatingly:

$$C(\Phi) = \sum_{(\nabla L, \nabla B)} \left(\lambda \|\Phi\|_\alpha^\alpha + \mu \|\Phi - \nabla B\|_2^2 \circ M + \gamma \|\Phi - \nabla L\|_2^2 \right) \quad (10)$$

$$C(L) = \sum_{(\nabla L, \nabla B)} \left(\omega \|\nabla L * K - \nabla B\|_2^2 + \gamma \|\Phi - \nabla L\|_2^2 \right) \quad (11)$$

To optimize L,there actually exists one close-form formulation for L as that in kernel estimation phase. Let $\partial C(L)/\partial L = 0$, and because of the convolution operation, Fourier transform is used, finally we can get the formulation:

$$L = \mathcal{F}^{-1} \left(\frac{\omega \mathcal{F}(\nabla) \circ \overline{\mathcal{F}(\nabla)} \circ \mathcal{F}(B) \circ \overline{\mathcal{F}(K)} + \gamma \mathcal{F}(\Phi) \circ \overline{\mathcal{F}(\nabla)}}{\omega \mathcal{F}(\nabla) \circ \overline{\mathcal{F}(\nabla)} \circ \mathcal{F}(K) \circ \overline{\mathcal{F}(K)} + \gamma \mathcal{F}(\nabla) \circ \overline{\mathcal{F}(\nabla)}} \right) \quad (12)$$

To optimize Φ, it's also a non-convex problem for $0 < \alpha < 1$. For given α and λ, μ, γ, only ∇B and ∇L are unknown. To use the lookup table trick in this case, the 2D table is built for various values of ∇B and ∇L. Because the two images are both normalized into 0-1 in advance, the range of these two derivatives are both from -1 to 1, the respective value of Φ in every pixel can be obtained by bilinear interpolation for every ∇B and ∇L with a fixed interval(0.0005 is chosen) in that range.

3 Experiments

In the experiment, we choose the images and kernels provided in [3] which has different kinds of kernels and includes the ground truth of the tested images. The chosen images and kernels are shown in Figure 3. What's more, some state-of-art models[2,4,9] in current are used to compare with our models.

Because of the kernel's tiny size, the fineness of the estimated kernels are not easy to be told by human's eyes.The method introduced in [10] will be taken as the method to evaluate the kernel-estimation quantitively.

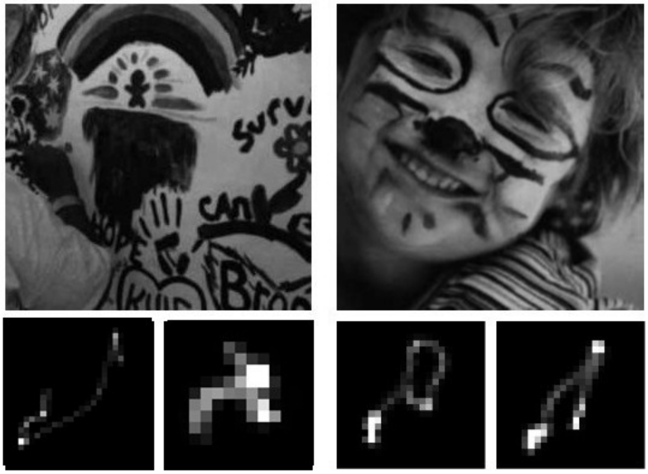


Fig. 3. The ground truth of the image and the kernel

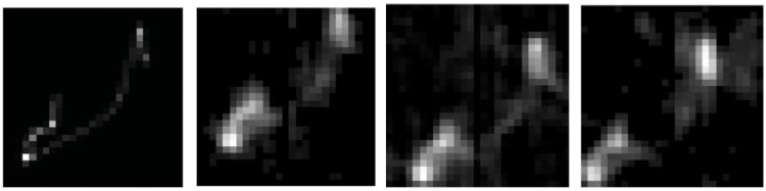
In this paper, we'll introduce a shift-PSNR method to get over the obstacle. Here we utilize the maximum cross-correlation value between estimated image and the ground-truth to find the shift coordinate:

$$S_\tau = max_\gamma \rho(G, \hat{L}, \gamma) \tag{13}$$

where G and L are ground truth and estimated image,γ is one possible shift of L. $\rho(\cdot)$ is defined as

$$\rho(G, L, \gamma) = \frac{\sum_\tau G(\tau) \cdot L(\tau + \gamma)}{\|G\|_2 \cdot \|L\|_2} \tag{14}$$

where τ is the corresponding coordinates. This equation $\rho(\cdot)$ is the normalized cross-correlation between G and L. After obtaining the shift coordinate, the PSNR can be calculated.



Origin kernel Super-sparse prior One-norm prior Two-norm prior

Fig. 4. Evaluation results of kernels with various priors

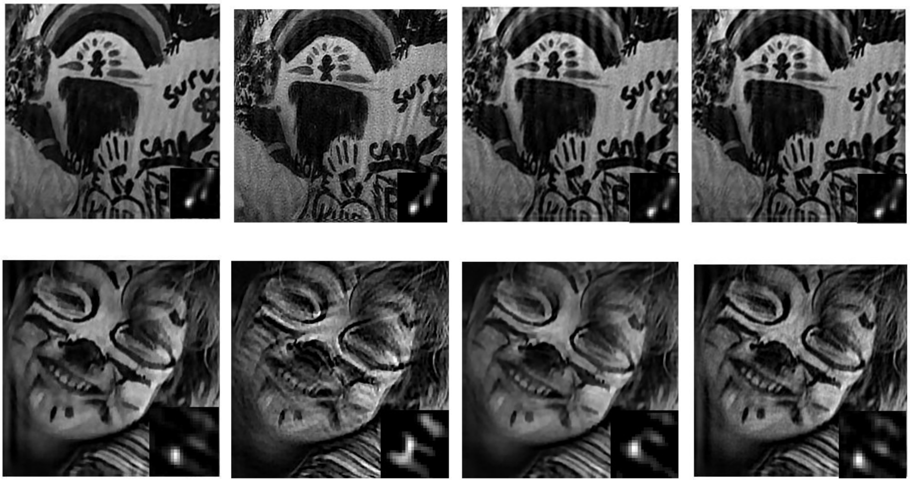
Table 1. Evaluation results:The similarity of the estimated kernels

Method	Img1,k1	Img1,k2	Img1,k3	Img1,k4	Img2,k1	Img2,k2	Img2,k3	Img2,k4
Our	**0.7334**	0.8110	0.8133	**0.8383**	**0.6146**	0.8039	**0.7121**	**0.7163**
Cho's	0.7060	**0.8582**	**0.8572**	0.8255	0.6033	**0.8149**	0.7037	0.5317
Shan's	0.6992	0.8072	0.7999	0.8111	0.5822	0.7438	0.5787	0.7061
Xu's	0.6683	0.7791	0.6705	0.5446	0.5881	0.7732	0.6880	0.5963

Table 2. Evaluation results:The shift-PSNR of the estimated image

Method	Img1,k1	Img1,k2	Img1,k3	Img1,k4	Img2,k1	Img2,k2	Img2,k3	Img2,k4
Our	23.6938	26.4998	**25.9776**	**24.3724**	**24.0196**	27.7135	**25.6521**	**24.5644**
Cho's	23.9201	**28.7057**	24.3193	22.6947	21.5643	27.2741	23.4266	17.3889
Shan's	22.5338	25.9742	24.6727	22.5818	20.5978	24.3099	20.2929	22.6934
Xu's	**24.5623**	26.0478	17.1736	17.7465	23.7963	**27.9993**	24.4097	18.5333

In Fig.4, super-sparsity prior can lead to a better estimated kernel obviously. In table 1 and 2, we can find that in most cases(5 of 8), our model of kernel estimation achieved better performance. The same result can also be found in Figure 4 in which we can find the estimated images with our model are more smooth and keep more details and the estimated kernels are more sparse.



Our result Shan's result Xu's result Cho's result

Fig. 5. Result of blind deconvolution

4 Conclusion

For the ill-posed motion deblurring problem, the priors on the images and kernels can disambiguate the redundancy solutions. In this paper, we propose a model with super-sparse prior during the kernel and latent image recovery. Considering the super-sparsity of the motion kernels and natural images, the priors will help to find the optimal solution faster and more accurate and avoid the solution to fall into suboptimal solution effectively. The experimental results also show that our model with super-sparse prior can reach more smooth and stable estimation.

Acknowledgement. The work was supported by the National Natural Science Foundation of China (Grant No. 61272251, 91120305).

References

1. Fergus, R., Singh, B., Hertzmann, A., Rowels, S.T., Freeman, W.T.: Removing camera shake from a single photograph. In: ACM SIGGRAPH, pp. 787–794 (2006)
2. Shan, Q., Jia, J., Agarwala, A.: High-quality motion deblurring from a single image. In: ACM SIGGRAPH, pp. 73:1–73:10 (2008)
3. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Understanding and evaluating blind deconvolution algorithm. In: CVPR, pp. 1964–1971 (2009)
4. Cho, S., Lee, S.: Fast motion deblurring. ACM Trans. Graph. 28 (2009)
5. Joshi, N., Szeliski, R., Kriegman, D.J.: PSF estimation using sharp edge prediction. In: CVPR (2007)
6. Kundur, D., Hatzinakos, D.: Blind image deconvolution. SPMag 13(3), 43–64 (1996)
7. You, Y.-L., Kaveh, M.: Blind image restoration by anisotropic regularization. IEEE Transactions on Image Processing 8, 396–407 (1999)
8. Yitzhaky, Y., Mor, I.: Direct method for restoration of motion-blurred images. Journal of Opt. Soc. Am. A 15(6), 1512–1519 (1998)
9. Xu, L., Jia, J.: Two-phase kernel estimation for robust motion deblurring. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 157–170. Springer, Heidelberg (2010)
10. Hu, Z., Yang, M.-H.: Good regions to deblur. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 59–72. Springer, Heidelberg (2012)
11. Bar, L., Sochen, N., Kiryati, N.: Image deblurring in the presence of salt-and-pepper noise. In: Kimmel, R., Sochen, N.A., Weickert, J. (eds.) Scale-Space 2005. LNCS, vol. 3459, pp. 107–118. Springer, Heidelberg (2005)
12. Wang, Y., Yang, J., Yin, W., Zhang, Y.: A new alternating minimization algorithm for total variation image reconstruction. SIAM J. Image Sciences 1(3), 248–272 (2008)

Proposal of Ultra-Short-Pulse Acoustic Imaging Using Complex-Valued Spatio-temporal Neural-Network Null-Steering

Kotaro Terabayashi and Akira Hirose

Department of Electrical Engineering and Information Systems, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
terabayashi@eis.t.u-tokyo.ac.jp, ahirose@ee.t.u-tokyo.ac.jp
http://www.eis.t.u-tokyo.ac.jp/

Abstract. We propose complex-valued spatio-temporal neural-network null-steering for wideband acoustic imaging with ultra-short pulses. We combine a complex-valued neural network (CVNN) and power-inversion adaptive array (PIAA) scheme to realize a practical-resolution imaging in the azimuth direction (direction perpendicular to the range direction) even with a small-aperture array. Simulations suggest that the proposed method presents a higher resolution than conventional methods such as Capon's method as well as a real-valued neural-network PIAA method.

Keywords: Acoustic imaging, power-inversion adaptive array (PIAA), complex-valued spatio-temporal neural network (CVSTNN).

1 Introduction

Acoustic imaging visualizes scattering caused by abrupt changes in acoustic impedance. It is often used under water and/or inside materials, in which optical techniques is unavailable, for applications in health diagnosis and non-destructive inspection. In usual human body imaging, a high frequency ultrasonic wave is transmitted from wide aperture sensor elements. Electronic switching of the elements realizes beam scanning for the imaging. This method cannot acquire images for a low-frequency low-loss scattering observation and/or with a use of small aperture sensor elements to be inserted into small space because of a large diffraction phenomenon.

Another scanning method is electronic adaptive beamforming using a sensor array. There exist several beamforming techniques. The most basic one is the delay and sum (DAS) method. The so-called Capon's method [1] is also a widely used beamforming and imaging technique to realize a sharp beam as a main lobe with suppression of side lobes [2**?** , 3]. In all the methods mentioned above, however, the steering angle depends on the wavelength of the acoustic wave. Then the frequency bandwidth of the transmitted wave has to be narrow enough to realize a fine alignment of the wavefront. A narrow band wave pulse possesses a long wave packet, and requires some special technique to obtain a high resolution in the range direction, i.e., direction along the propagation, for precise observation [**?**]. A narrow band wave can also cause breaking of target objects through acoustic resonance.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 217–224, 2013.

In this paper, we propose ultra short-pulse acoustic imaging using complex-valued spatio-temporal neural-network (CVSTNN) null-steering. A wideband system has low energy density in the frequency spectrum and, hence, it is less invasive. The short pulse leads to a higher range resolution and less speckle noise.

2 Complex-Valued Spatio-temporal Neural Network Power-Inversion Adaptive Array (CVSTNN-PIAA) and Its Use in Imaging

2.1 Imaging Procedure

We propose an acoustic imaging method consisting of a complex-valued spatio-temporal neural network (CVSTNN) [6] and the power-inversion adaptive array (PIAA) technique [7]. In this paper, we assume discrete target objects, and realize the imaging by steering nulls to the targets to estimate the DoA for imaging.

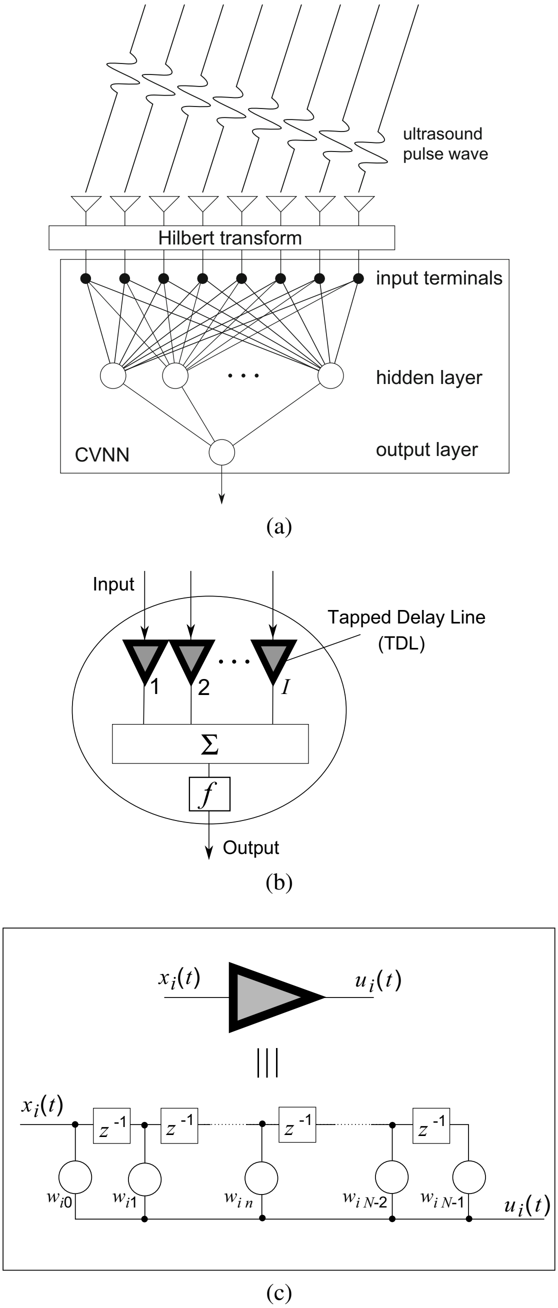
In our proposal, we conduct our imaging based on the PIAA. That is, in a certain acoustic field, we adjust the neural weights in the network in such a way that the signal is minimized at the neural output by using a learning process. Then we invert the profile of the directional sensitivity of the antenna system to obtain an image expressing the acoustic field. A set of learning iteration yields a single acoustic image. A change in the field requires another learning. However, the calculation cost is not very large if the field changes gradually, for example, if a target moves continuously. This is because we can use the weights for a previous field as their initial state in the following learning to reduce the calculation time since the acoustic field changes only slightly, resulting in only a small variation in the neural weights.

Let us consider M discrete urtrasonic scatterers (targets) in space. Ultrasound is transmitted to the space, scattered at the targets, and received at the sensor array. Each sensor generates time-sequential M pulses at maximum. We estimate the DoA based on the obtained time-sequential signals for the multiple scatterers to generate a scatterers' image.

We train the CVSTNN, or just CVNN, to direct its nulls to the estimated DoA of the M pulses. The learning is realized by feeding to the CVNN the received signals as teacher input signals while constant zeros as a teacher output signal. Then the nulls are directed to the pulse arrival directions. Next we calculate the radiation pattern (directional sensitivity) of the array by generating, feeding and steering a formal pulse numerically within the CVNN. The peaks in the inverse of the radiation pattern show the estimation of the target directions, which is equivalent to a scatterer space image.

2.2 Construction

For applications in mobile communications, we previously presented a wideband beamforming method using complex-valued layered neural networks [6]. There we proposed a CVNN for spatio-temporal complex-valued signal processing. This section presents its construction and the learning process to be used in the imaging system mentioned in the previous section.



(a)

(b)

(c)

Fig. 1. Complex-valued neural network to realize beamforming of ultra wideband signals (a) basic construction of CVNN, (b) neuron structure, and (c) tapped delay line (TDL)

Fig.1 shows the construction of the CVNN consisting of an input terminal layer, a hidden neuron layer and an output neuron layer. Signals detected at the sensors are converted into analytical signals by Hilbert transform which makes the phase values of positive- and negative-frequency components advance and retardant by $\pi/2$ rad, respectively. The analytical signal is expressed by the Hilbert transform $\mathcal{H}$ and received signals $\boldsymbol{x}$ as $(1+j\mathcal{H})\,\boldsymbol{x}$. It is fed to and processed by the CVNN.

Fig.1(b) illustrates the inside of a neuron. The signals run through tapped delay lines (TDLs) which work as synaptic weights. Then they are summed up to become the neural internal state u, and finally to generate a neural output in the range of $[-1,1]$ as $f(u)=\tanh(|u|)\exp(i\arg(u))$ where $f(u)$ is an activation function.

Fig.1(c) presents the structure of the TDL working as a synaptic weight. The z^{-1} box stands for a unit delay, and $w_{in}=a_{in}\exp(j\theta_{in})$ is a weight at the n-th tap for i-th input. The output of the TDL is the weighted sum of the time-sequential signals existing in the delay line, and generate the internal state u, as

$$u_i(t)=\sum_{n=0}^{N-1} x_i(t-n)w_{in} \tag{1}$$

$$u(t)=\sum_{i} u_i(t) \tag{2}$$

The CVNN realizes an adaptive beamformer for wideband signals in combination with the sensor array by adjusting the amplitude and phase values of the TDL weights $w(n)$ through complex-valued neural learning mentioned below.

2.3 Learning Dynamics of Complex-Valued Spatio-temporal Neural Network (CVSTNN)

The learning dynamics is explained as follows [8]. The outline is shown in Fig.2. First we prepare a set of teacher signals as combinations of input signals and corresponding desired output signals for the supervised learning. The input signals propagate forward to generate output, whereas the teacher output signals propagate backward to the input. In this learning, the teacher signals themselves, instead of errors, backpropagate through the network. The teacher signals $\hat{y}^{l-1}$ in layer $l-1$ is given by the teacher signals $\hat{\boldsymbol{y}}^{l}=[\hat{y}_j^l]^T\equiv[|\hat{y}_j^l|\exp\{j\hat{\theta}_j^l\}]^T$ and weights $\mathbf{W}^l=[\boldsymbol{w}_1^l\ \ldots\ \boldsymbol{w}_j^l\ \ldots\ \boldsymbol{w}_J^l]$ for $\boldsymbol{w}_j^l\equiv[w_{jin}^l]$ at n-th tap i-th input of j-th neuron in layer l as

$$(\hat{\boldsymbol{y}}^{l-1})^*=f(\,(\hat{\boldsymbol{y}}^{l})^*\,\mathbf{V}^{\mathbf{l}}\,) \tag{3}$$

where $\mathbf{V}^l=[v_{jin}]^l\equiv w_{jin}^l/|w_{jin}^l|^2$. Each neuron calculates the difference of the temporary output and the backpropagating teacher signal, and changes the weights according to the difference as [**?**]

$$\begin{aligned}|w_{jin}|^{\mathrm{new}}=|w_{jin}|^{\mathrm{old}}-K\Big\{\,&(1-|y_j|^2)\ \Big(|y_j|-|\hat{y_j}|\cos\Big(\theta_j-\hat{\theta}_j\Big)\Big)\,|x_{in}|\cos\theta_{jin}^{\mathrm{rot}}\\&-|y_j||\hat{y_j}|\sin\Big(\theta_j-\hat{\theta}_j\Big)\,\frac{|x_{in}|}{|u_j|}\sin\theta_{jin}^{\mathrm{rot}}\Big\}\end{aligned} \tag{4}$$

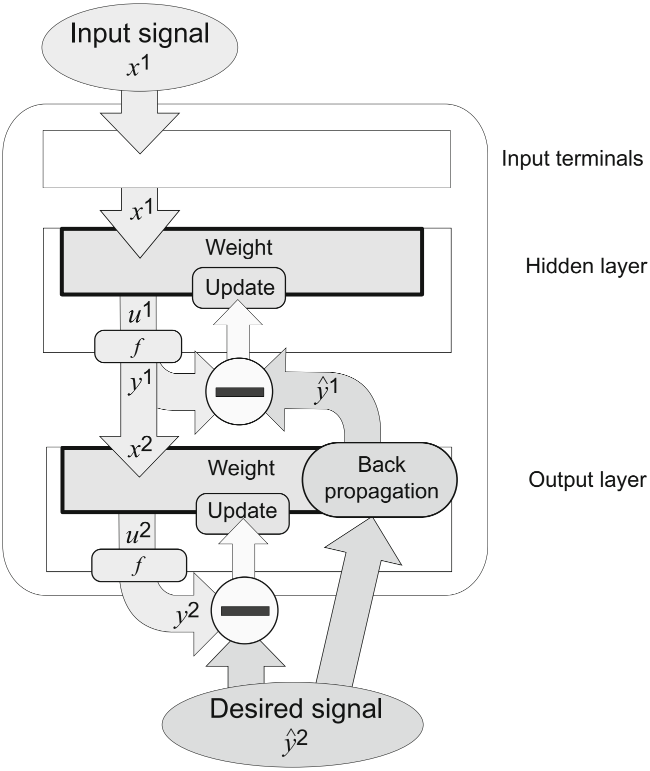


Fig. 2. Complex-valued backpropagation learning

$$\theta_{jin}^{\mathrm{new}} = \theta_{jin}^{\mathrm{old}} - K\Big\{ \left(1-|y_j|^2\right) \ \left(|y_j| - |\hat{y_j}|\cos\left(\theta_j - \hat{\theta_j}\right)\right)|x_{in}|\sin\theta_{jin}^{\mathrm{rot}} \\ +|y_j||\hat{y_j}|\sin\left(\theta_j - \hat{\theta_j}\right)\frac{|x_{in}|}{|u_j|}\cos\theta_{jin}^{\mathrm{rot}}\Big\} \qquad (5)$$

where $w_{jin} \equiv |w_{jin}|\exp\{\theta_{jin}\}$ (layer index l is omitted), $(\cdot)^{\mathrm{new}}$ and $(\cdot)^{\mathrm{old}}$ stand for update from old to new, $\theta_{jin} \equiv \arg(w_{jin})$, $\theta_j \equiv \arg(y_j)$ and $\theta_{jin}^{\mathrm{rot}} \equiv \theta_j - \theta_i - \theta_{jin}$.

In the iteration of the above learning process, the CVNN changes the output in such a way that the temporary output $\boldsymbol{y}^l$ converges at the desired one $\hat{\boldsymbol{y}}^l$ with appropriate (phase–amplitude focused) generalization characteristics arising from the complex-valued learning dynamics. The adaptive antenna system in total learns nulls for a set of teacher signals, namely, teacher input signals of incident pulses from arbitrary directions ϕ and an output signal of zero (no output), or learns to direct a beam to arbitrary directions ϕ with a pulse output teacher signal.

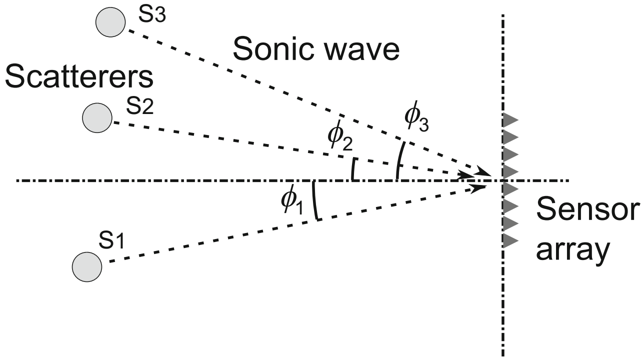


Fig. 3. Arrangement of sensor array and scatterers in simulation

3 Simulation Evaluation of Imaging Performance

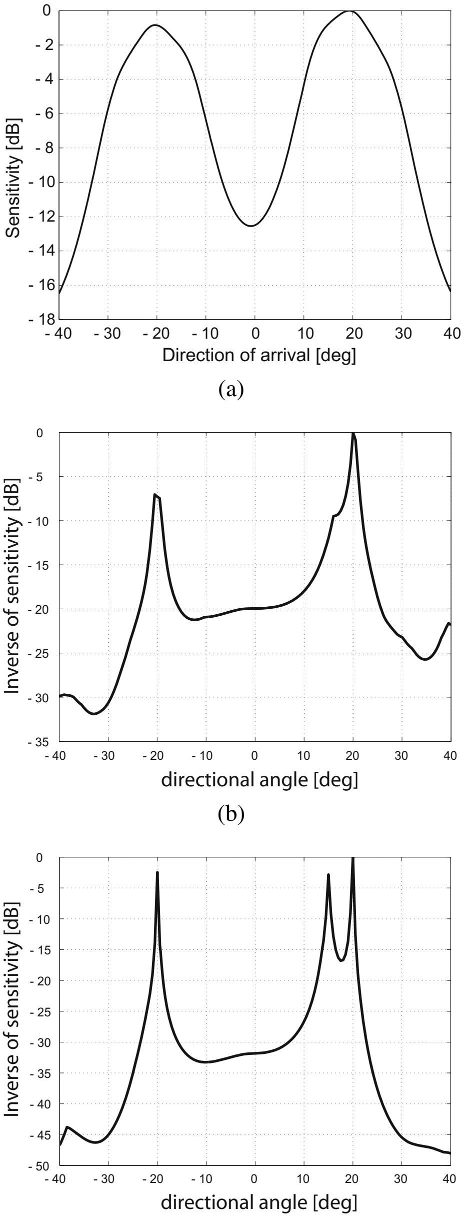
3.1 Experimental Setup

We evaluate the resolution of our proposed CVNN imaging method in a simulation. We estimate DoA in two-dimensional space by assuming a linear sensor array. We compare the results of conventional Capon's method, a real-valued neural network (RVNN) and the CVNN. The construction of the RVNN is the same as that of the CVNN including TDLs as the synaptic weights, but except for all the weights are real-valued, and the received signals are not converted to analytical signals. Both the CVNN and RVNN have 8 input terminals and 12 hidden neurons. Each TDL has 29 taps of 10^{-7} s unit delay.

Simulation parameters are chosen as follows. We assume three scatterers in directions of $\phi_1 = 0$, $\phi_2 = 10$, $\phi_3 = 15$, center frequency of 500 kHz, sampling frequency of 10 MHz, 8 ultrasonic sensors in a line with array pitch of 2 mm. Then the array size is very small, i.e., 14 mm. With this smallness, it is usually very difficult to obtain a high resolution in the azimuth direction. A linear array of eight sensors and three sactterers $S_m (m = 1, 2$ and 3) are positioned as shown in Fig. 3. When an ultrasonic short pulse is radiated on the space, it is scattered and received at the sensor array to generate three pulses at maximum at respective sensors. The time-sequential signals are fed to the neural network as a set of teacher signals, as well as a zero output teacher signal, to generate nulls to the scatterer directions. The pulse is an almost single duration pulse having a fractional bandwidth of 0.74.

3.2 Results

Fig.4 shows the result of one-dimensional imaging. The reference level is adjusted in such a way that the peaks of all the curves indicate 0 dB. The result of the Capon's method shows two gentle hills at around the scatterer angles of 0 deg, 10 deg and 15 deg, but the target location is not clear. The RVNN result presents two peaks, though 15 deg



(a)

(b)

(c)

Fig. 4. Acoustic imaging results for (a)conventional Capon's method, (b)RVNN-PIAA method, and (c)CVNN-PIAA method (proposal)

target cannot be separate from that of 20 deg. However, the peaks in the CVNN result are so sharp that we can distinguish the 15-deg and 20-deg targets. We find that the CVNN method has a very high resolution in the azimuth direction.

4 Summary

We proposed a wideband acoustic imaging method using a complex-valued spatio-temporal neural network. We combined the CVNN and PIAA to realize a high-resolution null learning even for very short pulses and small-aperture array. The simulation results suggested that the proposed CVNN-PIAA method shows a higher resolution than the conventional Capon's or the RVNN-PIAA method.

References

1. Capon, J.: High-resolution frequency-wavenumber spectrum analysis. Proceedings of the IEEE 57(8), 1408–1418 (1969)
2. Wang, Z., Li, J., Stoica, P., Nishida, T., Sheplak, M.: Constant-beamwidth and constant-powerwidth wideband robust capon beamformers for acoustic imaging. Journal of the Acoustical Society of America 116, 1621–1631 (2004)
3. Holfort, I., Gran, F., Jensen, J.: P2b-12 minimum variance beamforming for high frame-rate ultrasound imaging. In: IEEE Ultrasonics Symposium, pp. 1541–1544 (October 2007)
4. Synnevag, J.-F., Austeng, A., Holm, S.: Benefits of minimum-variance beamforming in medical ultrasound imaging. IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control 56(9), 1868–1879 (2009)
5. Nishino, T., Yamaki, R., Hirose, A.: Ultrasonic imaging for boundary shape generation by phase unwrapping with singular-point elimination based on complex-valued Markov random field model. IEICE Transactions on Fundamentals E93-A(1), 219–226 (2010)
6. Suksmono, A.B., Hirose, A.: Beamforming of ultra-wideband pulses by a complex-valued spatio-temporal multilayer neural network. International Journal of Neural Systems 15(1), 1–7 (2005)
7. Compton, R.: The power-inversion adaptive array: Concept and performance. IEEE Transactions on Aerospace and Electronic Systems AES-15(6), 803–814 (1979)
8. Hirose, A.: Applications of complex-valued neural networks to coherent optical computing using phase-sensitive detection scheme. Information Sciences –Applications– 2, 103–117 (1994)
9. Hirose, A., Yoshida, S.: Generalization characteristics of complex-valued feedforward neural networks in relation to signal coherence. IEEE Transactions on Neural Networks and Learning Systems 23, 541–551 (2012)

Computational Study of Depth Perception for an Ambiguous Image Region: How Can We Estimate the Depth of Black or White Paper?

Eiichi Mitsukura and Shunji Satoh

Graduate School of Information Systems, The University of Electro-communications, Tokyo, 182-8585 Japan
eiichi@hi.is.uec.ac.jp, shun@is.uec.ac.jp

Abstract. We propose a new computational model that accounts for human perception of depth for "ambiguous regions," in which no information exists to estimate binocular disparity as seen in black and white papers. Random dot stereograms are widely used examples because these patterns provide sufficient information for disparity calculation. Then, a simple question confronts us: "how can we estimate the depth of non-textured images, like those on white paper?" In such non-textured regions, mathematical solutions of the spatial disparities are not unique but indefinite. We examine a mathematical description of depth estimation that is consistent with psychological experiments for non-textured images. Using computer simulation, we show that resultant depth-maps using our model based on the mathematical description above qualitatively reproduce human depth perception.

Keywords: depth perception, depth propagation, binocular disparity, Gaussian curvature, ambiguous region, computational model.

1 Introduction

Our visual system estimates the depth and the slant information of objects from binocular images obtained by the right and left eye. Spatial disparities of the two images are determined by finding matching points. For example, as shown in Fig. 1**a**, the matching point of the upper left right angle in the "Left" image is the obtuse angle at the left-upper point of "Right" trapezoid. The spatial difference is referred to as binocular disparity. Unique solutions of horizontal disparities by finding matching points are limited along the right and left line segments of the rectangle or trapezoid (see the two lines in Fig. 1**b**). That is, no unique solution of disparity exists at any point in the black region. Hereinafter, we refer to such black regions as ambiguous regions.

Here, a simple question confronts us: "how do we estimate the depth of an ambiguous region?" Completion using determined disparity or depth (Fig. 1**b**) is one solution to have a unique value of disparity in the ambiguous region. Georgeson and his colleagues investigated how humans perceived the depth for ambiguous regions using psychological experiments [1]. They concluded that humans can realize depth completion by spatial propagation from the determined region into ambiguous regions.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 225–232, 2013.

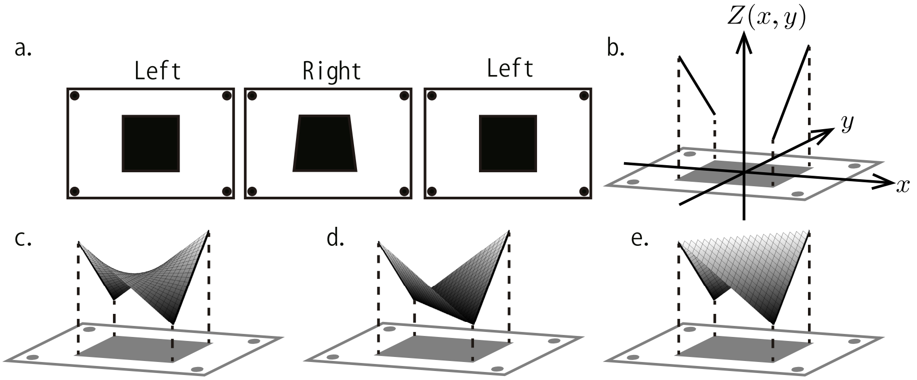


Fig. 1. a. Stereogram used in psychological experiments in [4]. The two left panels are parallel view methods; the two right panels are the cross-eyed view method. **b.** Two slanted lines show the depth calculated by binocular disparity in **a**. $Z(x,y)$ is the depth at a point (x,y). **c.** Example of depth propagation using a simple diffusion technique; hyperbolic paraboloid (saddle). **d** and **e** are typical results of human perception; flat depth maps.

Figure 1**c** portrays one example of depth propagation by the isotropic-diffusion generating as "smooth" a surface as possible [2]. Many computational models for stereopsis have used this kind of "smoothness" function (energy) defined by the first-order spatial derivative of the depth surface.

However, humans do not seem to perceive depth as shown in Fig. 1**c**. Ishikawa and Geiger showed that human perception differs from the image in Fig. 1**c**, but humans tend to recognize a "flat" surface as presented in Figs. 1**d** and 1**e**[3, 4]. Similar results were obtained for the case of Fig. 2**a**. Observing Fig. 1**d** and Fig. 2**d** (Fig. 1**e** and Fig. 2**e**), Ishikawa and Geiger found a common mathematical property of perceived depth: zero of the Gaussian curvature ($K=0$) [4]. Ishikawa obtained those flat surfaces so that resultant surfaces minimize an evaluation function of $|K|$ by simulated annealing, which is not a propagation scheme [3].

Our study specifically examines development of a psychologically acceptable model that completes the depth in the ambiguous region to $K=0$ by propagation of depth. Completion of visual information is also done in blind spots, where no visual information exists. We notice that the problem on depth completion is qualitatively equivalent to the problem on image completion at the blind spots (BS). Satoh and Usui propose a physiologically plausible model of the filling-in process of the BS [5]. Their model completes a lack of images at the BS by propagation of neural activities. Computationally, their neural network model minimizes curvature information of two kinds: (i) curvature of isophoto line $\bar{\kappa}^2$ and (ii) curvature of flow line $\bar{\mu}^2$. We can expect a demanding model for depth completion if we discover a close relation between K and $\{\bar{\kappa},\bar{\mu}\}$. We will be able to apply the BS model to the depth completion model.

This article is organized as follows. In Section 2, we present our proposed model for depth completion. In Section 3, we show a numerical simulation of our model. Section 4 presents conclusions.

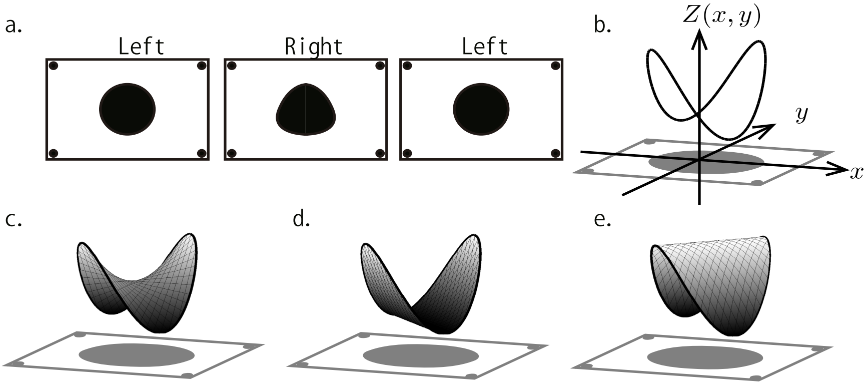


Fig. 2. a. Another example used in psychological experiments in [4]. **b.** The close curve displays the depth obtained from binocular disparity. **c.** An example of depth propagation (saddle) using a simple diffusion technique. **d** and **e** are typical results of human perception (flat) .

2 Proposed Model: Depth Completion by Propagation

2.1 Image Completion by Filling-in Model

A brief explanation of the filling-in model [5] is provided. Based on the computational theory of the filling-in model, we will derive a mathematical formula (model) of our model for depth completion.

The brightness I to be filled-in is signified by $I(x,y)$ defined in the (x,y) coordinate system. To evaluate the "goodness" of completed images at the blind-spots, Satoh and Usui defined an evaluation function $E[I]$ using curvature information $\bar{\kappa}^2$ and $\bar{\mu}^2$ as

$$E[I] = \iint_B (\bar{\kappa}^2 + \bar{\mu}^2)(I_x^2 + I_y^2)dxdy, \tag{1}$$

where $\frac{\partial}{\partial x}I(x,y) \equiv I_x$, $\frac{\partial}{\partial y}I(x,y) \equiv I_y$. Similarly, the second-order partial derivatives are denoted by I_{xx}, I_{xy} and I_{yy}. The two kinds of curvature information (see Fig. 3 for detail) are given as

$$\bar{\kappa}(x,y) = \frac{I_y^2 I_{xx} - 2I_x I_y I_{xy} + I_x^2 I_{yy}}{I_x^2 + I_y^2}, \tag{2}$$

$$\bar{\mu}(x,y) = \frac{(I_x^2 - I_y^2)I_{xy} - I_x I_y (I_{yy} - I_{xx})}{I_x^2 + I_y^2} . \tag{3}$$

The evaluation function E reaches the minimum value if

$$(\bar{\kappa}^2 + \bar{\mu}^2)(I_x^2 + I_y^2) = 0 \text{ for all } (x,y) \in \mathrm{B}. \tag{4}$$

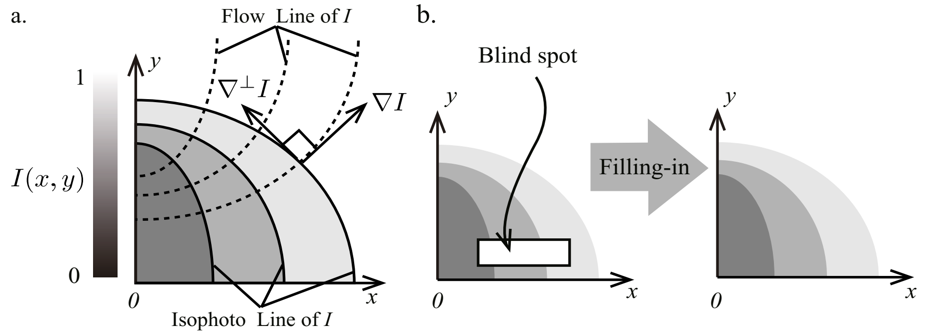


Fig. 3. Schematic explanation of the filling-in model [5] **a.** Solid curves illustrate isophoto lines (contour of the brightness). Dashed curves are referred to as flow curves, which are perpendicular to isophoto lines. The gradient vector ∇I gives the direction of the largest spatial change of brightness. And $\nabla^{\perp} I$ is perpendicular to ∇I. **b.** The region of blind-spot of left image is given as a white rectangle. Applying filling-in processing, resultant images have small curvatures of isophoto and flow lines.

2.2 Depth Completion Problem involves Image Completion Problem

We can expect that the evaluation function $E[I]$ is applicable for our depth completion model because $E[I] = 0$ if the surface $I(x,y)$ is "flat" $K = 0$ as shown in Figs. 1**d**–1**e** and Figs. 2**d**–2**e**. We prove our expectation mathematically as

$$\iint_{\mathrm{B}} (\bar{\kappa}^2 + \bar{\mu}^2)\left(I_x^2 + I_y^2\right) dxdy = 0 \Rightarrow \iint_{\mathrm{B}} K^2\, dxdy = 0, \tag{5}$$

where the Gaussian curvature is formulated as

$$K(x,y) = \frac{I_{xx}I_{yy} - I_{xy}^2}{\left(1 + I_x^2 + I_y^2\right)^2}\ . \tag{6}$$

Details are given in the Appendix (A) section. Now we can replace the brightness $I(x,y)$ with depth information $Z(x,y)$. Moreover, we introduce a time variable t and $Z(x,y,t)$ to represent the iterative update of Z.

Our model for depth completion by an iterative method which decreases $E[Z]$ as time progresses is written with the following dynamics:

$$\frac{\partial}{\partial t} Z = \nabla(\Delta Z) \cdot \nabla^{\perp} Z + \bar{\kappa}\,|\nabla Z|\,, \tag{7}$$

where $\nabla^{\perp} Z$ is perpendicular to ∇Z. Equation (7) propagates depth information Z because it is a kind of convection–diffusion equation. Equation (7) is equivalent to the dynamics used in a digital image inpainting problem (DII) [6, 7]

We can derive a dynamics for depth completion by the steepest descent algorithm to decrease $\iint \mathrm{K}^2 dxdy$. However, the resultant equation is composed of 129 terms (see Appendix (B) section), which is too complicated to obtain an expected results.

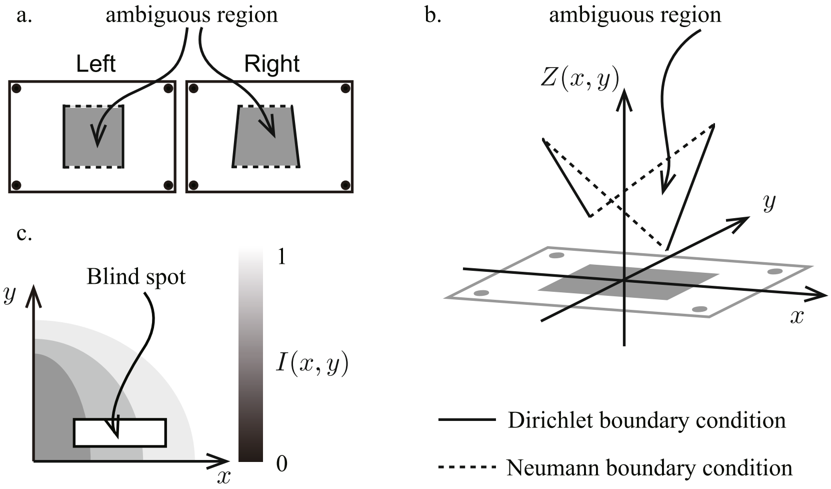


Fig. 4. The difference of boundary conditions between depth completion and image completion problem. **a**. Boundary condition on depth completion. The depths of right and left line segments are fixed as determined by horizontal disparities, whereas the depth of upper and bottom line segments are free from pre-determined values. **b**. Mathematically, those boundary conditions are referred to respectively as Dirichlet boundary conditions. **c**. On the filling-in problem at the blind-spot and DII problem, all boundaries are Dirichlet conditions.

2.3 Boundary Condition of Depth Propagation

The boundary condition of our model is different from the filling-in and DII problem, although the updating rule of depth propagation eq. (7) is equal to that of filling-in and DII problems.

On the depth completion problem (Figs. 4**a** and 4**b**), the depth around right and left line segments must be fixed during depth updating by eq. (7). This restriction is referred to mathematically as the Dirichlet boundary condition of differential equations. The depth of the upper and bottom line segments can be updated during depth completion by iterative update, *i.e.* the Neumann condition. Around the region to be updated with the Neumann condition, depth update or diffusion is restricted to the horizontal direction x.

3 Numerical Simulation

We evaluate whether the completed depths by our model are consistent with human perceptual results by numerical simulation of eq. (7). The results obtained using our model are presented in Fig. 5. Figures 5**a** and 5**b** are the steady states of eq. (7) starting with determined boundaries presented in Fig. 1**b**. Both results are consistent with human perception as shown in Figs. 1**d** and 1**e**; “flat” surfaces. The model presented herein also generates two solutions: a concave surface and a convex flat surface. The differences of solutions are attributable to the different initial values of depth in the ambiguous region. Figure 5**a** was obtained by initial value

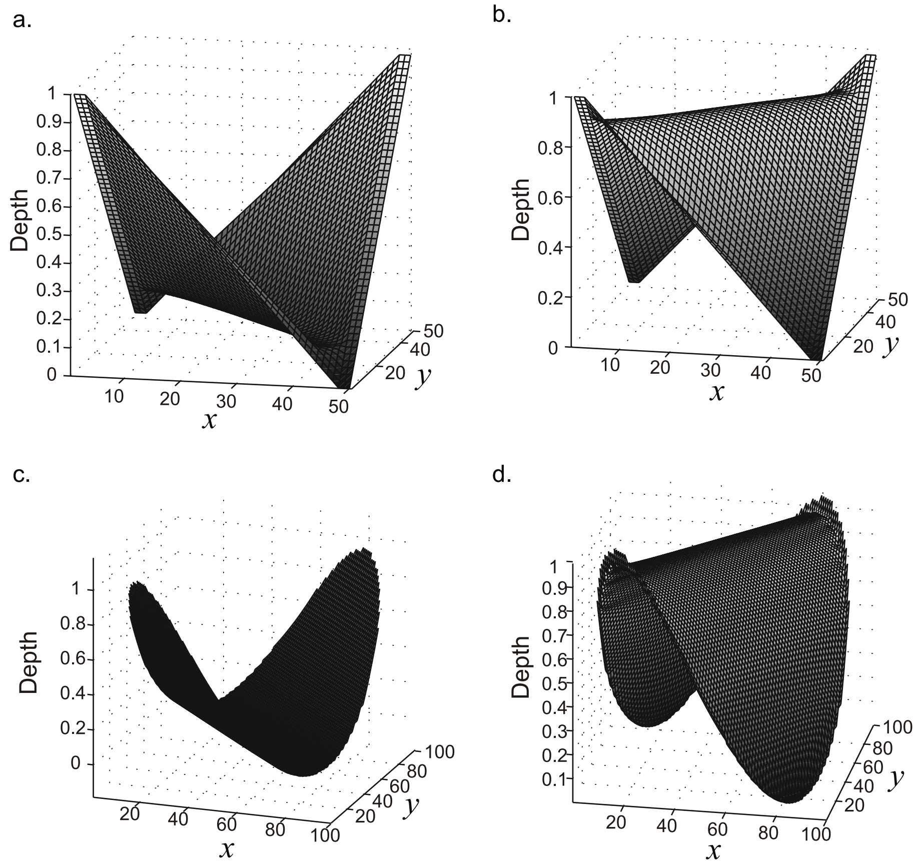


Fig. 5. Completed depth by the proposed model. These results are caused by different initial conditions. Panels **c** and **d** show results of numerical simulation to Fig. 2. These results depend on initial conditions as in panels **a** and **b**.

$Z(x, y, t = 0) = 0.0$, whereas Figure 5**b** came from $Z(x, y, t = 0) = 1.0$. Similar results were obtained for Fig. 2**b** as displayed in Figs. 5**c** and 5**d**.

4 Conclusion

We propose a computational model for depth completion by a propagation scheme. Numerical experimental results were consistent with human perception. That is, completed depth values in the ambiguous region were "flat" $K = 0$. Moreover, two solutions (concave or convex surfaces) were generated using our model. Our model might reproduce bi-stable perception of depth in case of ambiguous regions.

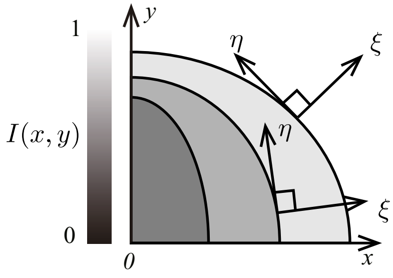


Fig. 6. (ξ,η) defines local coordinate systems at each spatial position of $I(x,y)$

Our model requires *a priori* information about ambiguous/not-ambiguous region. For example, the right and left edges in Fig.1a give unique disparity information, while white background and black box are ambiguous regions. Propagation direction of depth should be restricted into the black-box. This condition is too simplified to apply the model for natural images.

Future works include application of our model to natural images by automatic detection of ambiguous region, and examination of physiological evidence supporting our computational model.

Appendix (A)

We prove the proposition of eq. (5). $\bar{\kappa}$, $\bar{\mu}$ and K are rotation invariant. Therefore, we can calculate them on the local coordinate system (ξ,η) so that $I_\eta = 0$ (see Fig. 6). The direction ξ is parallel with ∇I; η is perpendicular to ξ. We can locally rotate $I(x,y)$ so that $I_y = 0$ for all (x,y). Note that rotation angles differ at each position (x,y). Because of rotation invariant $\bar{\kappa}$, $\bar{\mu}$, and K, such local rotation does not affect the proposition of eq. (5). For such reasons, we can assume $I_y = 0$ for all (x,y).

If the sufficient condition is satisfied, then $\bar{\kappa} = \bar{\mu} = 0$ or $I_x = 0$ for all (x,y). (Remember $I_y = 0$) If $I_x \neq 0$, then $I_{yy} = 0$ and $I_{xy} = 0$ because $\bar{\kappa} = 0$ and $\bar{\mu} = 0$. Substitution $I_{yy} = I_{xy} = 0$ for eq. (6), $K = 0$. However, if $I_x = 0$, then $K \neq 0$.

Q.E.D.

Appendix (B)

The update rule of decreasing $\iint K^2 dxdy$ by the steepest descent method is the following complex equation composed of

$$\begin{aligned}\frac{\partial Z(x,y,t)}{\partial t} = \big(&-4Z_{yy}\,Z_{xy}^4 + 36Z_y^2 Z_{yy} Z_{xy}^4 - 4Z_{yy} Z_x^2 Z_{xy}^4 + 80 Z_y Z_x Z_{xy}^5 \\ &- 32 Z_y Z_{xy}^3 Z_{xyy} - 32 Z_y^3 Z_{xy}^3 Z_{xyy} - 32 Z_y Z_x^2 Z_{xy}^3 Z_{xyy} + 8 Z_{yy}^2 Z_{xy}^2 Z_{xx} \\ &- 72 Z_y^2 Z_{yy}^2 Z_{xy}^2 Z_{xx} + \cdots + (\mathbf{117\ terms}) + 2 Z_{yy}^2 Z_x^2 Z_{xxxx} \\ &+ 2 Z_y^2 Z_{yy}^2 Z_x^2 Z_{xxxx} + Z_{yy}^2 Z_x^4 Z_{xxxx}\big) \,/\, \big(1 + Z_x^2 + Z_y^2\big)^6.\end{aligned}$$

Acknowledgement. Author E.M would like to thank Mr.Tetsumasa Asano for his helpful comments and suggestions. This work was partially supported by JSPS KAKENHI Grant Number 50333844.

References

1. Mark, A.G., Yates, T.A., Schofield, A.J.: Depth propagation and surface construction in 3-D vision. Vision Research 49, 84–95 (2009)
2. Belhumeur, P.N.: A Bayesian approach to binocular stereopsis. International Journal of Computer Vision 19, 237–262 (1996)
3. Ishikawa, H.: Total Absolute Gaussian Curvature for Stereo Prior. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 537–548. Springer, Heidelberg (2007)
4. Ishikawa, H., Geiger, D.: Illusory volumes in human stereo perception. Vision Research 46(1-2), 171–178 (2006)
5. Satoh, S., Usui, S.: Computational theory and applications of a filling-in process at the blind spot. Neural Networks 21, 1261–1271 (2008)
6. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques – SIGGRAPH 2000, pp. 417–424 (2000)
7. Satoh, S.: Computational identity between digital image inpainting and filling-in process at the blind spot. Neural Computing and Applications 21, 613–621 (2011)

Learning a Sparse Representation for Robust Face Recognition

Weihua Ou[1,2], Xinge You[1], Pengyue Zhang[1], Xiubao Jiang[1], Ziqi Zhu[1],
and Duanquan Xu[1]

[1] Department of Electronics and Information Engineering,
Huazhong University of Science and Technology, Wuhan 430074, China
{ouweihuahust,penryzhang,jiangxiubao}@gmail.com,
youxg@mail.hust.edu.cn, xuduanquan@126.com, ziqi_zhu@163.com
[2] Department of mathematics, Huaihua University, Huaihua 418008, China

Abstract. Based on the assumption that occlusions have sparse representation on the nature pixel coordinate, Sparse Representation based Classification (SRC) [9] adopts an identity matrix as occlusion dictionary to deal with the occlusions or noises. However, this assumption is often violated in real applications, such as the faces are occluded by scarf. In this paper, we present an approach to learn an occlusion dictionary from the data. Thus, the occlusions have sparse representation on the learned occlusion dictionary and can be effectively separated from the occluded face images. Experimental results show our approach achieves better performance than SRC, while the computational cost is much lower.

Keywords: Face recognition, occlusions, dictionary learning, learning sparse representation.

1 Introduction

Though current face recognition techniques have reached a certain level of maturity in controlled settings, the complex intra-class variations, such as pose, illumination, expression and occlusions, are difficult to model and lead to recognition failures. Among them, occlusions are one of the most challenging problems. In real applications, the use of accessories, such as sunglasses, scarves, hats, or objects placed in front of the face can be viewed as occlusions. Moreover, violations of an assumed model for face appearance may act like occlusions, e.g., shadows due to extreme illumination violate the assumption of a low-dimensional linear illumination model [4]. Many methods are proposed to deal with occlusions, such as, localized non-negative matrix factorization [6], Local Binary Patterns [2] and Gabor wavelets [7]; however, these methods only operate on the non-occluded regions, and aim to circumvent the occluded regions, rather than to recover the occluded parts of face image, which might be essential for recognition.

Unlike the above methods, Sparse Representation based on Classification (SRC) [9] proposed by Wright et al. aims to eliminate the occlusions. Based on the assumption that occlusions have sparse representation on nature pixel coordinate, SRC adopts an identity matrix as the occlusion dictionary, and seeks the sparse representation over the

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 233–240, 2013.

expanded dictionary which consists of the training sample dictionary and the occlusion dictionary. If the sparse representation is recovered correctly, the occlusion will be effectively eliminated and classification can be performed based on the reconstruction error only using the sparse representation coefficients over the training sample dictionary. The experiments shows that SRC has achieved the best performance for random noises. However, the experiments on the AR database also shows that, SRC is not nearly as robust to contiguous occlusion as it is to random pixel corruption. For sunglasses and scarf, it achieves only 87% and 59.5% respectively. The main reasons of this is that the assumption is violated and the representation is not sparse.

In this paper, we learn an occlusion dictionary from data to obtain the sparse representation and conduct recognition based on the learned structural sparse representation, which we call SSRC. First, we model the occluded faces as the summation of the non-occluded faces and the occlusions. Then we present a fast algorithm to learn an occlusion dictionary from data. Compared to the identity matrix occlusion dictionary used in SRC, the occlusions can be sparsely represented by the prototype of occlusion atoms. At the same time, the size of the expanded dictionary is significantly reduced with respect to the identity matrix occlusion dictionary in SRC. This will accelerate the speed of sparse coding for each test sample. Fig. 1 presents the illustration of the proposed approach. The lower images in Fig. 1(b) and 1(c) show the learned occlusion dictionary and the coefficients on this learned occlusion dictionary. It can clearly be seen that the atoms of the learned occlusion dictionary closely resemble the occlusion by scarf, and the coefficients are very sparse. As shown in the top image of Fig. 1(d), the occlusions are successfully separated and the recovered non-occluded face image is perfect except for the edge of the occlusion. Thus, the classification can be efficiently conducted on the recovered non-occluded face image.

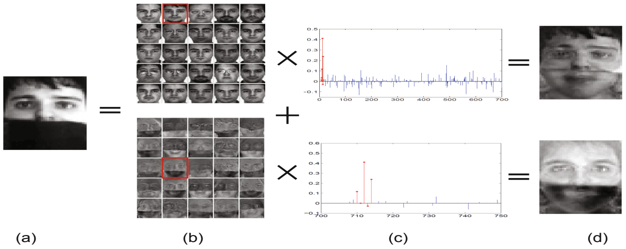


Fig. 1. Illustration of SSRC: an occluded face image (a) is represented as a sparse linear combination of the training sample dictionary and the occlusion dictionary in (b). The decomposed sparse coefficients in (c) correspond to the dictionary. The non-occluded face image and the occlusion in (d) can be jointly estimated.

The remainder of the paper is organized as follows. In Section 2, we describe the occlusion dictionary learning algorithm. In Section 3, we present the recognition algorithm based on the learned sparse representation. Finally, we conduct experiments in Section 4 and conclude this paper in Section 5.

2 Learning Sparse Representation

2.1 The Model of Occlusions

We denote the training samples of the i-th class as $\mathbf{A}_i = [\boldsymbol{s}_{i,1}, \boldsymbol{s}_{i,2}, \cdots, \boldsymbol{s}_{i,p_i}] \in \mathbb{R}^{m \times p_i}$ and each column of the matrix $\mathbf{A}_i$ denotes a sample. Suppose we have k classes and gather the training samples of all classes to build the training sample dictionary $\mathbf{A}_0 = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_k] \in \mathbb{R}^{m \times p}$, where $p = p_1 + p_2 + \cdots + p_k$ is the number of training samples and p_i is the training sample number of i-th class. For the occluded face $\boldsymbol{y}$, we model it as follows:

$$\boldsymbol{y} = \boldsymbol{y}_0 + \boldsymbol{y}_d + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$, $\boldsymbol{y}_0$, $\boldsymbol{y}_d$ denotes the occluded face, the non-occluded face, and the occlusions, respectively, $\boldsymbol{e}$ is an error term that compensates the noise. Based on the assumption that the training samples from a single class lie on a subspace, $\boldsymbol{y}_0$ can be represented sparsely over the training sample dictionary $\mathbf{A}_0$, i.e., $\boldsymbol{y}_0 = \mathbf{A}_0\boldsymbol{\alpha}_0$, where $\boldsymbol{\alpha}_0$ is the sparse representation coefficients. In the next subsection, we present an approach to learn an occlusion dictionary, which can sparsely represent $\boldsymbol{y}_d$.

2.2 Learning Occlusion Dictionary

Given a data set $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n] \in \mathbb{R}^{m \times n}$ for occlusion dictionary learning, we first project them onto the corresponding class and utilize the associated projection residuals $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n] \in \mathbb{R}^{m \times n}$ to train the occlusion dictionary. For each sample $\boldsymbol{y}_r, r = 1, 2, \cdots, n$, suppose it belongs to class i, the projection residual is $\boldsymbol{p}_r = \boldsymbol{y}_r - \mathbf{A}_i(\mathbf{A}_i^T\mathbf{A}_i)^{-1}(\mathbf{A}_i^T\boldsymbol{y}_r)$, where $\mathbf{A}_i$ is the training sample of class i in $\mathbf{A}_0$.

We denote the occlusion dictionary as $\mathbf{A}_d = [\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_q] \in \mathbb{R}^{m \times q}$, where each atom is of unit length, i.e., $\boldsymbol{d}_j^T\boldsymbol{d}_j = 1, j = 1, 2, \cdots, q$. According to above discussions, we expect that $\mathbf{A}_d$ can sparsely represent $\boldsymbol{y}_d$ associated with the occlusion part, and at the same time "bad" for the non-occluded face image $\boldsymbol{y}_0$. We formulate the objective function for learning occlusion dictionary as follows:

$$\begin{aligned} &\min_{\mathbf{A}_d, \mathbf{\Lambda}} \|\mathbf{P} - \mathbf{A}_d\mathbf{\Lambda}\|_F^2 + \lambda_1\|\mathbf{\Lambda}\|_1 + \lambda_2\|\mathbf{A}_0^T\mathbf{A}_d\|_F^2 \\ &s.t. \quad \boldsymbol{d}_j^T\boldsymbol{d}_j = 1, \quad j = 1, 2, \cdots, q, \end{aligned} \tag{2}$$

where $\mathbf{A}_0$ is the training sample dictionary, $\mathbf{\Lambda} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n] \in \mathbb{R}^{q \times n}$ contains sparse coefficients of the projection residuals $\mathbf{P}$ on dictionary $\mathbf{A}_d$, λ_1 and λ_2 are the regularization parameters. The regularization term $\|\mathbf{A}_0^T\mathbf{A}_d\|_F^2$ encourages the incoherence between $\mathbf{A}_0$ and $\mathbf{A}_d$, and thus the learned occlusion dictionary $\mathbf{A}_d$ prefers to be as independent as possible to the training sample dictionary $\mathbf{A}_0$. As a result, the non-occluded face image can be efficiently obtained by seeking the sparse representation of the test face image on the expanded dictionary, i.e., $\mathbf{A} = [\mathbf{A}_0, \mathbf{A}_d]$.

Formulation.(2) is a joint optimization problem of the occlusion dictionary $\mathbf{A}_d$ and sparse coefficients $\mathbf{\Lambda}$. Although it is not jointly convex on both, it is convex on $\mathbf{A}_d$ (or $\mathbf{\Lambda}$) given fixed $\mathbf{\Lambda}$ (or $\mathbf{A}_d$). Similar to dictionary learning algorithm [1], we optimize

$\mathbf{A}_d$ and $\mathbf{\Lambda}$ alternatively, i.e., we optimize $\mathbf{A}_d$ given fixed $\mathbf{\Lambda}$, and then optimize $\mathbf{\Lambda}$ given fixed $\mathbf{A}_d$. The two steps are conducted iteratively until convergence. The whole algorithm is shown below.

Algorithm 1. Occlusion Dictionary Learning

Input: Training data $\mathbf{Y}$, training sample dictionary $\mathbf{A}_0$, λ_1, λ_2
Output: The occlusion dictionary $\mathbf{A}_d$ and the sparse coefficients $\mathbf{\Lambda}$
Initialization:We initialize each column of $\mathbf{A}_d$ as a random vector with unit l_2-norm
Step 1: Compute the projection residuals $\mathbf{P}$
Step 2: Fix $\mathbf{A}_d$ and optimize $\mathbf{\Lambda}$

$$\min_{\mathbf{\Lambda}} \|\mathbf{P} - \mathbf{A}_d\mathbf{\Lambda}\|_F^2 + \lambda_1\|\mathbf{\Lambda}\|_1 \tag{3}$$

Step 3: Fix $\mathbf{\Lambda}$ and optimize $\mathbf{A}_d$. We update each atom $\boldsymbol{d}_l$ of the dictionary $\mathbf{A}_d$ separately with all the other atoms $\boldsymbol{d}_{j\neq l}$ fixed, sweep through the columns and always use the most updated atoms that emerge from the preceding step. The update rule is given by

$$\begin{aligned} \boldsymbol{d}_l &= \left[(\boldsymbol{\beta}_l\boldsymbol{\beta}_l^T - \gamma)\mathbf{I} + \lambda_2\mathbf{A}_0\mathbf{A}_0^T\right]^{-1}\mathbf{Z}\boldsymbol{\beta}_l^T \\ \boldsymbol{d}_l &= \boldsymbol{d}_l/\|\boldsymbol{d}_l\|_2 \end{aligned} \tag{4}$$

Conduct steps 2 and 3 iteratively until the maximum number of iterations is reached or the values of the adjacent objective functions are sufficiently close.

3 Face Recognition via the Learned Sparse Representation

By concatenating the learned occlusion dictionary with the training sample dictionary, we obtain a structured dictionary $\mathbf{A} = [\mathbf{A}_0, \mathbf{A}_d]$. Given a test sample$\boldsymbol{y}$, we formulate the structured sparse recovery problem as follows:

$$\{\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_d\} = \arg\min_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_d} \|\boldsymbol{y} - \mathbf{A}_0\boldsymbol{\alpha}_0 - \mathbf{A}_d\boldsymbol{\alpha}_d\|_2^2 + \xi_1\|\boldsymbol{\alpha}_0\|_1 + \xi_2\|\boldsymbol{\alpha}_d\|_1. \tag{5}$$

We use the l_1-l_s [5] to solve it. Once the sparse solution is computed, denote the recovered face as $\hat{\boldsymbol{y}}_0 = \boldsymbol{y} - \mathbf{A}_d\hat{\boldsymbol{\alpha}}_d$, then the reconstruction error is computed with respect to the recovered face as follows:

$$r_i(\boldsymbol{y}) = \|\hat{\boldsymbol{y}}_0 - \mathbf{A}_0\delta_i(\hat{\boldsymbol{\alpha}}_0)\|_2, \quad i = 1, 2, \cdots, k. \tag{6}$$

Finally, the identity is the class corresponding to the minimum reconstruction error. The whole procedure is presented in algorithm 2.

Algorithm 2. Structured Sparse Representation based Classification (SSRC)

```
**Input**: Test sample $\boldsymbol{y} \in \mathbb{R}^m$, $\xi_1$, $\xi_2$
**Output**: Identity of test sample $\boldsymbol{y}$
**Initialization**: Training sample dictionary $\mathbf{A}_0$ and the learned occlusion dictionary $\mathbf{A}_d$
**Step 1**: Compute the sparse representation of $\boldsymbol{y}$ on the structured dictionary $\mathbf{A}$
    $\{\hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_d\} = \arg\min_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_d} \|\boldsymbol{y} - \mathbf{A}_0\boldsymbol{\alpha}_0 - \mathbf{A}_d\boldsymbol{\alpha}_d\|_2^2 + \xi_1\|\boldsymbol{\alpha}_0\|_1 + \xi_2\|\boldsymbol{\alpha}_d\|_1$
**Step 2**: Compute the residuals
    $r_i(\boldsymbol{y}) = \|\boldsymbol{y} - \mathbf{A}_d\hat{\boldsymbol{\alpha}}_d - \mathbf{A}_0\delta_i(\hat{\boldsymbol{\alpha}}_0)\|_2, \quad i = 1, 2, \cdots, k \qquad (7)$
**Step 3**: Output the identity: Identity$(\boldsymbol{y}) = \arg\min_i r_i(\boldsymbol{y})$.
```

4 Experiments

In this section, we conduct experiments to evaluate the performance of SSRC. We compare to following algorithms: (1) sparse representation based classification (SRC) [9], in which an identity matrix is utilized as the occlusion dictionary; (2) Gabor feature based sparse representation for face recognition (GSRC) [8], in which a Gabor occlusion dictionary is learned; (3) robust sparse coding for face recognition (RSC) [10], which needs several iterations of sparse coding for each test sample; and (4) Extended SRC: undersampled face recognition via intra-class variant dictionary(ESRC) [3], in which an intra-class variant dictionary is constructed by subtracting the class centroid of images from the same class. For SSRC, we learn one occlusion dictionary with the mutual incoherence regularization term and the other occlusion dictionary without. We denote them as SSRC1 and SSRC2, respectively.

4.1 Datasets and Parameter Setting

For SRC and ESRC, we implement the error-constrained model with the same error tolerance used in the original paper [9,3], i.e., $\varepsilon = 0.05$. For RSC and GSRC, we use the programs provided by the authors[1]. Since SSRC performed stably for wide ranges of model parameters, we set $\lambda_1 = 0.02$, $\lambda_2 = 0.5$, and $\xi_1 = \xi_2 = 0.001$. For SSRC, the occlusion dictionary size q is set to 50.

We conduct experiments on the AR database by choosing a subset with 50 men and 50 women. We resize each image into different resolutions 12×10, 26×20, 31×24, 42×30, and 51×40, which correspond to feature dimensions 120, 520, 744, 1,260 and 2,040, respectively. We consider following three scenarios.

Sunglasses. In this scenario, one image with sunglasses in session one for each of the 100 subjects is randomly chosen as a training sample for training the occlusion dictionary, while the test set consists of seven images without occlusion in session two and

[1] `http://www4.comp.polyu.edu.hk/~cslzhang/code.htm`

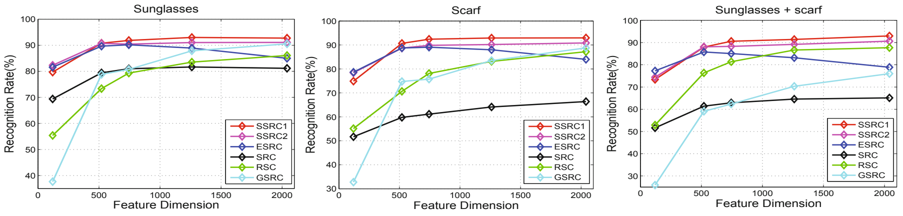
the remaining images with sunglasses in both sessions for each subject. In total, we have twelve test images for each subject.

Scarf. Similar to the sunglasses scenario, one image with scarf in session one for each of the 100 subjects is randomly chosen for training the occlusion dictionary, and the remaining images with scarf in both sessions and the seven images without occlusion in session two for each subject are used for testing.

Sunglasses+Scarf. The third scenario is that two occluded images (one with sunglasses and one with scarf) for each of the 100 subjects are randomly chosen from the session one are used for training the occlusion dictionary, and the rest seventeen images for each subject are used for testing. The challenge of this case is that there are two kinds of disguise with illumination variations, expression variations and whether the faces are occluded or not is unknown to the algorithm. In all three scenarios, the training sample dictionary is the same and consists of seven images without occlusion in session one.

4.2 Recognition Rate

The recognition results are shown in Fig.2. It shows that SSRC1 and SSRC2 greatly outperform SRC, RSC and GSRC. In these three scenarios, SSRC1 achieves the maximal recognition rate at 92.99%, 92.74%, 92.59% and outperforms SRC by 12.01%, 26.35%, and 27.47%, respectively. Both SRC and GSRC have a much worse recognition rate in dealing with "Sunglasses + Scarf", while RSC performs better than SRC and GSRC in almost all dimensions.

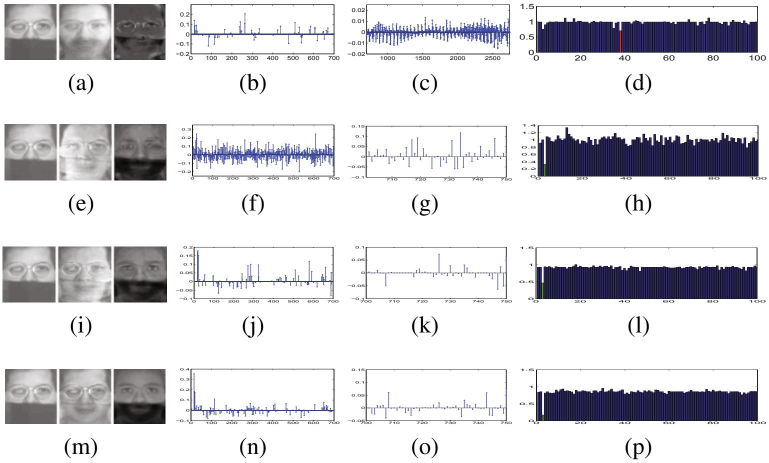


(a) (b) (c)

Fig. 2. Recognition rates of different methods on a subset of the AR database with the feature dimensions varying from 120 to 2,040: (a) Sunglasses, (b) Scarf, (c) Sunglasses +scarf

4.3 Comparison for Representation Coefficients

The assumption in SRC is that the occlusions can be sparsely coded by the identity matrix occlusion dictionary. This assumption might be violated when there are severe occlusions. Fig. 3 shows such an example, in which roughly 40% of the whole face is occluded by the scarves. It is obvious that the occlusion of scarves can not be sparsely represented by the identity matrix occlusion dictionary as shown in Fig. 3(c). The reconstructed face using SRC is completely different to the original face, as presented in the middle subfigure of Fig. 3(a); severe shadows are found in the mouth area and the outline of the forehead is severely distorted, which influence classification correctness. As presented in Fig. 3(d), the red bar indicates that SRC fails to identify the subject. However, our learned occlusion dictionary represents the scarf occlusion sparsely as shown in Fig. 3(o), many of the coefficients are zeros. Obviously, the reconstructed face image using SSRC1 is almost perfect except for a light scarf mark which does not obscure the true identity. The green bar in Fig. 3(p) shows the correct class. Despite ESRC and SSRC2 identify the subject correctly, the reconstruction residuals is larger than that of SSRC1, which means that they are less robust than SSRC1.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

Fig. 3. Comparison between SRC, ESRC, SSRC1 and SSRC2 on the AR database with scarf: the upper row is the SRC result, the second row is the ESRC result, the third row is the SSRC2 result, and the lower row is the SSRC1 result. (a)(e)(i)(m) A test face image from subject 3 in the AR database: the left, middle and right subfigures correspond to the occluded image, reconstructed image and reconstruction error image, respectively. (b)(f)(j)(n) Sparse coefficients associated with the training sample dictionary. (c)(g)(k)(o) Sparse coefficients associated with the occlusion dictionary. (d)(h)(l)(p) Reconstruction residuals with respect to the coefficients for different classes; the green bar indicates the correct class.

5 Conclusion

In this paper, we present an approach to learn a sparse representation for robust face recognition. A occlusion dictionary is learned by mutual incoherence regularization. The learned occlusions dictionary can sparsely represent the occluded parts of faces. Apart from the improved recognition rate, an important advantage of SSRC is its compact occlusion dictionary, which has many fewer atoms than used in SRC [9]. This greatly speed the sparse coding. We evaluate the proposed method on different scenarios, including extreme variations of illumination, expressions and real disguises. The experimental results clearly demonstrate the proposed method achieves better performance than existing sparse representation methods.

Acknowledgements. This work was supported partially by the National Natural Science Foundation of China (Grant No.61272203), the International Scientific and Technological Cooperation Project (Grant No. 2011DFA12180), National Science & Technology R&D Program (Grant No.2012BAK31G01, Grant No. 2012BAK02B06), the Ph.D Programs Foundation of Ministry of Education of China (Grant No. 20110142110060), and the Key Science and Technology Program of Wuhan (Grant No.201210121021).

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing 54(11), 4311–4322 (2006)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on PAMI 28(12), 2037–2041 (2006)
3. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. IEEE Transactions on PAMI 34(9), 1864–1870 (2012)
4. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on PAMI 23(6), 643–660 (2001)
5. Kim, S., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: An interior-point method for large-scale l1-regularized least squares. IEEE Journal of Selected Topics in Signal Processing 1(4), 606–617 (2007)
6. Lee, D., Seung, H., et al.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
7. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing 11(4), 467–476 (2002)
8. Yang, M., Zhang, L.: Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 448–461. Springer, Heidelberg (2010)
9. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on PAMI 31(2), 210–227 (2009)
10. Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: Proc. of CVPR, pp. 625–632. IEEE (2011)

Stacked Denoising Autoencoders for Face Pose Normalization

Yoonseop Kang, Kang-Tae Lee, Jihyun Eun,
Sung Eun Park, and Seungjin Choi

[1] Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 790-784, Korea
[2] KT Advanced Institute of Technology
17 Woomyeon-dong, Seocho-gu, Seoul, Korea
{e0en,seungjin}@postech.ac.kr,
{kangtae.lee,eunjihyun,sungeun.park}@kt.com

Abstract. The performance of face recognition systems are significantly degraded by the pose variations of face images. In this paper, a global pose normalization method is proposed for pose-invariant face recognition. The proposed method uses a deep network to convert non-frontal face images into frontal face images. Unlike existing part-based methods that require complex appearence models or multiple face part detectors, the proposed method relies only on a face detector. The experimental results using the Georgia tech face database demonstrate the advantages of the proposed method.

Keywords: Pose normalization, face recognition, autoencoder, stacked denoising autoencoder.

1 Introduction

Unlike traditional face recognition systems that recognize large number of people, mobile devices or televisions need to recognize only a small set of people with high accuracy. Although there are large amount of benchmark datasets available, collecting enough amount of training images of the set of people of interest is still very difficult. Therefore, we need a face recognition system that guarantees high accuracy with a small amount of training data.

When given a small number of training data, it is hard to generalize over the many kinds of variations including pose and illumination. While effect of changes in illumination can be reduced by using preprocessing techniques including histogram normalization, the changes in pose is difficult to handle with simple pre-processing. One of the popular approaches to overcome the difficulty of generalization over pose changes is to use pose normalization on face images.

Pose normalization refers to methods that infers a frontal face when given a non-frontal face images. Most of pose normalization algorithms are part-based: They locate face parts, and re-locate them to their standard positions [1][2]. The

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 241–248, 2013.

main drawback of the part-based methods is that they require all face parts to be located with high accuracy. Locating face parts is done by sophisticated models including AAMs [3], and these models are not suitable for mobile devices with limited processor speed and memory size.

Instead of part-based methods, we suggest a global method that only uses a single detector: a face detector. The proposed method takes a whole non-frontal face image and converts it into a frontal face image. To learn the complex mapping between non-frontal and frontal faces, the proposed method uses a deep network called stacked denoising autoencoder.

In this paper, we first review the stacked denoising autoencoders, and then describe our framework for pose normalization and face recognition. Experiments on a benchmark dataset shows the usefulness of the proposed method in improving face recognition accuracy.

2 Stacked Denoising Autoencoder

2.1 Denoising Autoencoder

The term *autoencoder* refers to an unsupervised, deterministic neural network that 1) generates a hidden representation from input, 2) and then reconstructs input from the hidden representation [4]. Therefore, an autoencoder is composed of two parts: an encoding function and a decoding function (Fig. 1).

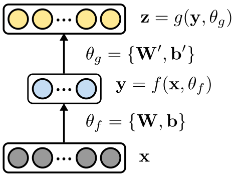

$\mathbf{z} = g(\mathbf{y}, \theta_g)$
$\theta_g = \{\mathbf{W}', \mathbf{b}'\}$
$\mathbf{y} = f(\mathbf{x}, \theta_f)$
$\theta_f = \{\mathbf{W}, \mathbf{b}\}$
$\mathbf{x}$

Fig. 1. Structure of an autoencoder with encoding function $f(\boldsymbol{x}, \theta_f)$ and decoding function $g(\boldsymbol{y}, \theta_g)$

An encoding function $f(\boldsymbol{x}, \theta_f)$ maps an input $\boldsymbol{x}$ to a hidden representation $\boldsymbol{y}$ using a affine transformation with a projection matrix $\boldsymbol{W}$ and a bias $\boldsymbol{b}$, followed by a non-linear squashing function. Sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$ is typically used as a squashing function.

$$\boldsymbol{y} = f(\boldsymbol{x}, \theta_f) = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{1}$$

Then a decoding function $g(\boldsymbol{y}, \theta_g)$ maps the hidden representation back to a reconstruction of input $\boldsymbol{z}$. A decoding function can be either linear or nonlinear. Affine transformation is often used when the input takes real values, and sigmoid squashing function is applied when the input is binary:

$$\boldsymbol{z} = g(\boldsymbol{y}, \theta_g) = \begin{cases} \boldsymbol{W}'\boldsymbol{y} + \boldsymbol{b}' & \text{or} \\ \sigma(\boldsymbol{W}'\boldsymbol{y} + \boldsymbol{b}') \end{cases} \tag{2}$$

Training an autoencoder is done by minimizing the mean-squared reconstruction error with respect to parameters $\theta_f = \{\boldsymbol{W}, \boldsymbol{b}\}$ and $\theta_g = \{\boldsymbol{W}', \boldsymbol{b}'\}$.

$$\arg\min_{\theta_f, \theta_g} \mathbb{E}\{||\boldsymbol{x} - \boldsymbol{z}||_2^2\} \tag{3}$$

Although autoencoders learn an effective encodings that are capable of reconstructing inputs, they suffer from overfitting when the dimension of hidden representations becomes higher. Moreover, it is likely that such autoencoders to learn a trivial identity mapping, instead of learning useful features from data.

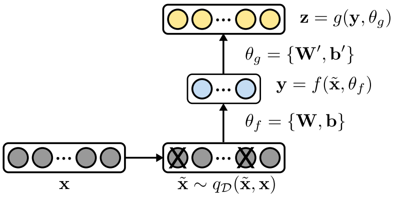

$\mathbf{z} = g(\mathbf{y}, \theta_g)$
$\theta_g = \{\mathbf{W}', \mathbf{b}'\}$
$\mathbf{y} = f(\tilde{\mathbf{x}}, \theta_f)$
$\theta_f = \{\mathbf{W}, \mathbf{b}\}$
$\mathbf{x}$
$\tilde{\mathbf{x}} \sim q_{\mathcal{D}}(\tilde{\mathbf{x}}, \mathbf{x})$

Fig. 2. Structure of an denoising autoencoder with encoding function $f(\boldsymbol{x}, \theta_f)$, decoding function $g(\boldsymbol{y}, \theta_g)$, and stochastic corruption $q_{\mathcal{D}}(\tilde{\boldsymbol{x}}|\boldsymbol{x})$

Denoising autoencoder (DAE) was proposed to overcome the limitations of autoencoders by reconstructing denoised inputs $\boldsymbol{x}$ from corrupted, noisy inputs $\tilde{\boldsymbol{x}}$ [5]. DAEs avoids overfitting and learns better, non-trivial features by introducing stochastic noises to training samples. One may generate corrupted inputs $\tilde{\boldsymbol{x}}$ from their original value $\boldsymbol{x}$ with several different stochastic corruption criteria $q_{\mathcal{D}}(\tilde{\boldsymbol{x}}|\boldsymbol{x})$, including adding Gaussian random noise, randomly masking dimensions to zero, and adding salt-pepper noise (Fig. 2).

$$\tilde{\boldsymbol{x}} \sim q_{\mathcal{D}}(\tilde{\boldsymbol{x}}|\boldsymbol{x}) \tag{4}$$
$$\boldsymbol{y} = f(\tilde{\boldsymbol{x}}, \theta_f) = \sigma(\boldsymbol{W}\tilde{\boldsymbol{x}} + \boldsymbol{b}) \tag{5}$$
$$\boldsymbol{z} = g(\boldsymbol{y}, \theta_g) = \boldsymbol{W}'\boldsymbol{y} + \boldsymbol{b}' \tag{6}$$

The objective function of DAEs remains the same as typical autoencoders. Note that the objective function minimizes the discrepancy between reconstructions and original, uncorrupted inputs $\boldsymbol{x}$, not the corrupted inputs $\tilde{\boldsymbol{x}}$.

A DAE is trained using back-propagation just as ordinary multi-layer perceptrons.

2.2 Stacked DAEs

Stacking DAEs on top of each other allows the model to learn more complex mapping from input to hidden representations [6]. Just as other deep models

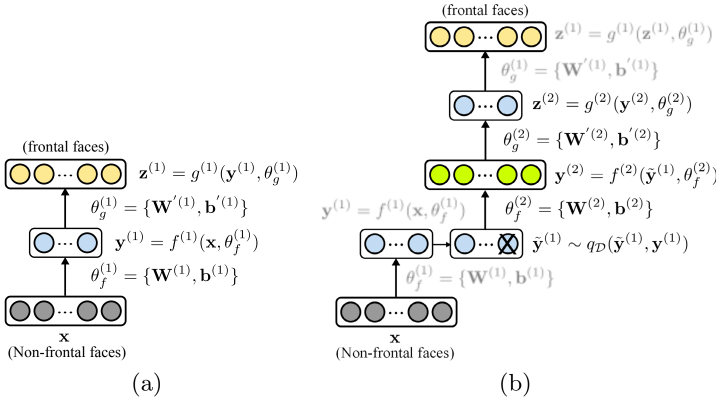

(frontal faces)
$\mathbf{z}^{(1)} = g^{(1)}(\mathbf{y}^{(1)}, \theta_g^{(1)})$
$\theta_g^{(1)} = \{\mathbf{W}'^{(1)}, \mathbf{b}'^{(1)}\}$
$\mathbf{y}^{(1)} = f^{(1)}(\mathbf{x}, \theta_f^{(1)})$
$\theta_f^{(1)} = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}\}$
x
(Non-frontal faces)
(a)
(frontal faces)
$\mathbf{z}^{(1)} = g^{(1)}(\mathbf{z}^{(1)}, \theta_g^{(1)})$
$\theta_g^{(1)} = \{\mathbf{W}'^{(1)}, \mathbf{b}'^{(1)}\}$
$\mathbf{z}^{(2)} = g^{(2)}(\mathbf{y}^{(2)}, \theta_g^{(2)})$
$\theta_g^{(2)} = \{\mathbf{W}'^{(2)}, \mathbf{b}'^{(2)}\}$
$\mathbf{y}^{(2)} = f^{(2)}(\tilde{\mathbf{y}}^{(1)}, \theta_f^{(2)})$
$\mathbf{y}^{(1)} = f^{(1)}(\mathbf{x}, \theta_f^{(1)})$
$\theta_f^{(2)} = \{\mathbf{W}^{(2)}, \mathbf{b}^{(2)}\}$
$\tilde{\mathbf{y}}^{(1)} \sim q_{\mathcal{D}}(\tilde{\mathbf{y}}^{(1)}, \mathbf{y}^{(1)})$
$\theta_f^{(1)} = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}\}$
x
(Non-frontal faces)
(b)

Fig. 3. Pre-training of stacked DAEs. (a) Train a bottom layer DAE with clean and corrupted inputs (in our case, frontal faces and non-frontal faces), then (b) train another DAE that reconstructs $\boldsymbol{z}^{(2)}$ from the hidden representation $\boldsymbol{y}^{(1)}$ extracted from the bottom layer DAE.

including deep belief networks [7], training stacked DAEs is also done in two-phase: layerwise, greedy pre-training and fine-tuning.

Unlike typical deep models that are extended by adding layers from bottom to top in pre-training, stacked DAEs are extended by adding layers in the middle of them. More specifically, the pre-training of stacked DAEs is done by the following steps.

First, train bottom layer DAE with encoding function $\boldsymbol{y}^{(1)} = f^{(1)}(\boldsymbol{x}, \theta_f^{(1)})$ and decoding function $\boldsymbol{z}^{(1)} = g^{(1)}(\boldsymbol{y}^{(1)}, \theta_g^{(1)})$ (Fig. 3(a)). Once the bottom layer DAE is trained, train a new DAE that takes the hidden representations of the bottom layer DAE $\boldsymbol{y}^{(1)}$ as training data. Stochastic noise $q_{\mathcal{D}}(\tilde{\boldsymbol{y}}^{(1)}|\boldsymbol{y}^{(1)})$ is added to $\boldsymbol{y}^{(1)}$ to generate corrupted input $\tilde{\boldsymbol{y}}^{(1)}$.

$$\boldsymbol{y}^{(1)} = f^{(1)}(\boldsymbol{x}, \theta_f^{(1)}) \tag{7}$$

$$\tilde{\boldsymbol{y}}^{(1)} \sim q_{\mathcal{D}}(\tilde{\boldsymbol{y}}^{(1)}|\boldsymbol{y}^{(1)}) \tag{8}$$

$$\boldsymbol{y}^{(2)} = f^{(2)}(\tilde{\boldsymbol{y}}^{(1)}, \theta_f^{(2)}) \tag{9}$$

$$\boldsymbol{z}^{(2)} = g^{(2)}(\tilde{\boldsymbol{y}}^{(2)}, \theta_g^{(2)}) \tag{10}$$

Train more DAEs in a similar way until the desired number of layers is achieved. After pre-training, the weights and biases of stacked DAE are fine-tuned by back-propagation as ordinary neural networks.

3 Pose Normalization Using Stacked DAEs

The number of face images collected from users is often insufficient to cover various changes in poses. On the other hand, it is relatively easy to obtain

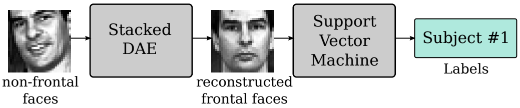


Fig. 4. The schematics of the proposed pose normalization and face recognition system composed of stacked DAEs and SVMs

large amount of face images with pose variations from the existing databases. Therefore, it is necessary to utilize the face databases to improve the performance of classifiers that are trained with relatively small number of images provided by users. Pose normalization methods can assist classifiers with the following procedure.

1. Train a pose normalization algorithm that learns a general mapping from non-frontal faces to frontal faces, using the existing face image databases.
2. Collect face images from users and train classifier with the collected images.
3. Given a non-frontal face image as a query, run the pose normalization on the query, then feed the pose-normalized image to classifier for recognition.

As mapping from non-frontal faces to frontal faces is highly complex, deep models like stacked DAEs are ideal choice for learning such mappings. Moreover, the fact that the learning procedure of DAEs is not affected by the type of corruption applied to inputs suggests that one can use more sophisticated procedures to corrupt inputs for DAEs, instead of just adding random noises to samples. Therefore, we consider non-frontal face images as corrupted versions of a frontal face, and learn mapping between non-frontal and frontal face images using stacked DAEs for pose normalization and face recognition.

As images take real-values, affine decoding function is used for the bottom layer of stacked DAE, and sigmoid function for the rest of layers. Sigmoid function was used for encoding function for all layers. As described above, Corrupted inputs for the bottom layer was given as non-frontal images, and inputs for higher layers were corrupted by Gaussian noises with standard deviation α (i.e. $q_{\mathcal{D}}(\tilde{\boldsymbol{y}}^{(1)}|\boldsymbol{y}^{(1)}) = \mathcal{N}(\boldsymbol{y}^{(1)}, \alpha^2\boldsymbol{I})$).

4 Face Recognition on Georgia Tech Face Database

4.1 Data Pre-processing

Georgia Tech face database [8] consists of pictures 50 subjects in 15 different poses. One of the 15 poses is frontal, and the variations among poses are relatively high (Fig. 5). We consider the frontal faces as uncorrupted inputs for stacked DAEs, and the remaining 14 non-frontal faces as corrupted inputs.

The resulting dataset is still too small to train a large deep network. Therefore, we applied additional corruptions as below on every frontal and non-frontal image to generate more corrupted samples:

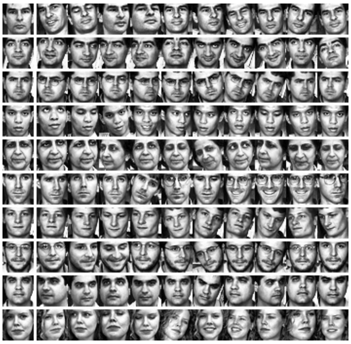


Fig. 5. Frontal and 10 non-frontal faces of 10 subjects from Georgia Tech face database

- Translate horizontally and vertically by up to 2 pixels.
- Rotate in -30 to 30 degrees.
- Flip horizontally.

By applying these corruption procedures, we expanded the original dataset with 750 images into a larger dataset with 26,550 images. All corrupted images were converted into grayscale, and histogram-normalized.

4.2 Experimental Settings

Two different experiment settings were tested to measure the performance of the proposed method.

1. Setting #1: To simulate the situation of having multiple training samples for each subject, whole dataset was randomly partitioned into training set and test set that contains 80% and 20% of the samples. Training set was used to train both stacked DAE and SVMs, and test set was used to test the proposed face recognition system.
2. Setting #2: To simulate an extreme case of having only single image for each subject, we used the face images of 80% of *subjects* for training stacked DAE, and used the remaining 20% as the training set for SVMs and test set for the proposed face recognition system.

For pose normalization, we trained 3-layer stacked DAE with 2000 and 1000 latent dimensions for each hidden layers (resulting into a network with 1024-2000-1000-2000-1024 nodes), and ran the stochastic gradient updates for 1,200 epochs with batch size 100. The noise level α was set to 0.2. We used linear support vector machine (SVM) and kernel SVM with RBF kernel to classify the faces. We also ran SVMs on the the raw non-frontal face images (without passing them through stacked DAE) as a baseline.

Table 1. Face recognition accuracies on Georgia Tech face database

settings	baseline(linear)	baseline(RBF)	proposed(linear)	proposed(RBF)
setting #1	0.108	0.098	**0.843**	0.806
setting #2	0.235	0.289	**0.454**	0.409

4.3 Experimental Results

Before quantitatively analyzing the face recognition results, we first visualized the weights of stacked DAE learned from the pairs of corrupted non-frontal faces and frontal faces. Most of the weights contained circular filters which captures the rotations of the faces (Fig. 6(a)).

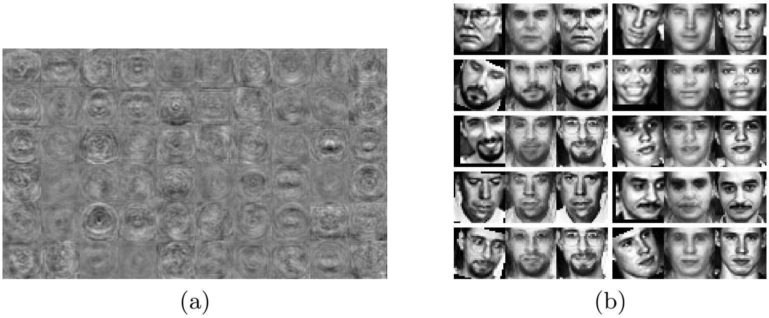


Fig. 6. (a) Subset of weights learned by the first layer of stacked DAE trained on Georgia Tech face database. Each small square corresponds to weight values between a hidden node and all input pixels. Higher pixel intensity indicates larger weight value. (b) 10 examples of corrupted face images from Georgia tech face database (left), their pose-normalized version obtained by the proposed method (middle), and ground-truth frontal face images (c).

The reconstructed frontal images obtained by processing non-frontal corrupted face images using stacked DAEs were a bit blurry, but still retained important characteristics of the ground-truth frontal face images (Fig. 6(b)).

The quantitative comparison reveals the effectiveness of the proposed approach more clearly. When relatively large number of training samples were provided (setting #1), the improvement of accuracy over naive baseline method was dramatic (84.3% vs. 10.8%). Even when only one training image was given (setting #2), the proposed method significantly improved the classification accuracy (45.4% vs. 23.5%).

5 Conclusion

In this paper, we introduced a pose normalization and face recognition system that uses stacked DAEs. By learning a general mapping from non-frontal faces to frontal faces using stacked DAEs, the proposed method performs pose normalization without any sophisticated appearence models. Experiments on Georgia Tech face database evaluated the effectiveness of the proposed system in terms of face recognition accuracy in usual settings and extreme settings with only single training sample per subject.

Acknowledgments. This work was supported by POSTECH & KT Open R & D Program.

References

1. Du, S., Ward, R.: Component-wise pose normalization for pose-invariant face recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 873–876 (2009)
2. Asthana, A., Marks, T.K., Jones, M.J., Tieu, K.H., Rohith, M.V.: Fully automatic pose-invariant face recognition via 3D pose normalization. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 937–944 (2011)
3. Cootes, T., Walker, K., Talyor, C.J.: View-based active appearance models. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 227–232 (2000)
4. Becker, S.: Unsupervised learning procedures for neural networks. The International Journal of Neural Systems 1 & 2, 17–33 (1991)
5. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 1096–1103 (2008)
6. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research 11, 3371–3408 (2010)
7. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)
8. Nefian, A.V.: Georgia tech face database (1999), `http://www.anefian.com/research/face_reco.htm`

A Wearable Cognitive Vision System for Navigation Assistance in Indoor Environment

Liyuan Li, Gang S. Wang, Weixun Goh, Joo-Hwee Lim, and Cheston Tan

Institute for Infocomm Research, Singapore, 138632
{lyli,gswang,wgoh,joohwee,cheston-tan}@i2r.a-star.edu.sg

Abstract. Existing mobile navigation techniques are not applicable for indoor navigation. Obviously, the best navigator is a human companion. In this paper, we explore to build a wearable virtual navigator for indoor navigation. A novel cognitive vision system is designed which consists of long-term memory and working memory for complicated vision tasks in dynamic environments. The long-term memory mimics the flexibility and scalability of human cognitive memory for domain knowledge representation, and the working memory emulates the routine process and attention selection in human cognitive model for online visual perception. Efficient algorithms for image classification and object detection are organized and performed under cognitive perception framework to achieve real-time performance. Field tests demonstrate its effectiveness and efficiency by recognizing scenes, locations, and landmark objects in real-time, and subsequently providing context-aware assistant to guide the user in the navigation of a complex office environment.

Keywords: Indoor navigation, cognitive architecture, long-term memory, working memory, scene recognition, object detection.

1 Introduction

Existing mobile navigation techniques are not applicable for indoor navigation due to the unavailability of GPS signals. An alternative chocie is using another wireless signals, such as RFID or WiFi signals, but the infrastructure cost may be too high for deployment [1]. Obviously, the best navigator is a human guider companying you in the journey, however, such a human guider is not available in most of the cases, even through remote teleoperation [2].

With the pervasiveness of portable computers and wearable cameras, it becomes possible to develop a wearable virtual guider based on visual perception for indoor navigation. Visual perception and context-awareness in changing environments underlies various cognitive skills and vision recognitions, such as recognition of scene, position, landmark or informative object, context-aware information searching and acquisition, and episodic reasoning. For such tasks, human cognitive vision system is considered much better than existing computer vision systems [3]. The objective of this paper is to explore the way to build a cognitive vision system for indoor navigation in a large and complex building. We

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 249–257, 2013.

propose a novel wearable cognitive vision system that combines scene categorization, location recognition, and landmark detection under a cognitive framework which consists of a hierarchical structure as a long-term memory for domain knowledge and a working memory for online computing. Efficient visual recognition methods are employed. The prototype system has been tested in real-world field and demonstrated its effectiveness and efficiency in helping with the process of navigating a typical office building.

1.1 Related Work

Indoor navigation of complex, unfamiliar environments is a challenging task. Industry solutions depend on various sensors. GPS-based navigation systems do not work indoors. Special sensors, *e.g.* RFID [1], may allow for accurate indoor localization, but the infrastructure costs (retro-fitting, etc.) may be prohibitive. A recently-developed system called Indoor Atlas [4] analyzes magnetic field perturbations inside a building to perform localization. These methods are map-based solutions [5,6], which require the system designers to input the detailed floorplans of a building, and then construct a mapping between the signals and the 2D locations on the map. Users are localized based on wireless signals. This solution provides a “You are Here” dot indication in a 2D map, but if users do not know where “here” is in relation to “there”, they can quickly become lost in a labyrinth of unknown twists and turns. In addition, wireless signals are often lost at rounding corners, along narrow, long corridors, and so on.

Existing approaches that rely on visual input for navigation are based on image-matching algorithms [7]. First, an image database containing multi-view images of all landmark locations is built. During navigation, the system compares the input with all the images in the database to find the best match. This approach works for small lab environments, but may not work for larger indoor environments for real-time navigation since over ten thousands of sample images may be required to form the database.

There are many cognitive architectures developed in cognitive science [8]. However, the purpose is to mimic the general and broad capabilities of human intelligence in learning and problem solving. Most of them do not address the challenges and opportunities specific to visual perception and memory [3]. Recently, a few advanced cognitive systems are developed for visual perception, but they focus on the simple object or action recognition from a fixed vision platform in a constrained environment [9,10]. They cannot be simply applied for visual navigation task in dynamic and complex environment.

2 The Cognitive Framework

Inspired by the progresses in cognitive architectures [8,3] and working memory [11], we propose a cognitive visual memory framework for a mobile agent. It consists of two main modules: a Long-Term Memory (LTM) model and a Working Memory (WM) or Short-Term Memory (STM) model. In long-term memory, we encompass a hierarchical structure to embed the spatial schema and visual appearance of the environment which involves context-related semantic, procedural, and episodic knowledge of visual perception for indoor navigation. The working memory includes visual features of input image, routine processes, attention selection, and actions for context-aware service.

2.1 Long-Term Memory Model

The domain knowledge regarding a complex office building is organized hierarchically according to the layers of cognitive concepts, as illustrated in Figure 1. The top level is the general description of the building, the second level contains scene descriptions, the next level contains distinctive locations in each scene, and the last level is for landmark or informative objects at specific locations. Each node in the hierarchy contains context-related information:

- Feature representation used for the recognition or detection, e.g., visual feature dictionary (BOW) for scene or location nodes or HOG descriptor for object nodes;
- Classifier or detector used for visual recognition or detection, e.g., a SVM classifier for scene or location recognition, or object detection;
- Spatial memory for efficient object detection and spatial-temporal reasoning;
- Semantic description of contextual information related to the node;
- Action for context-aware service, e.g., turn left in corridor.

When applying this system to a specific building, one can classify the indoor scenes into a few typical categories according to human concepts of indoor scenes,

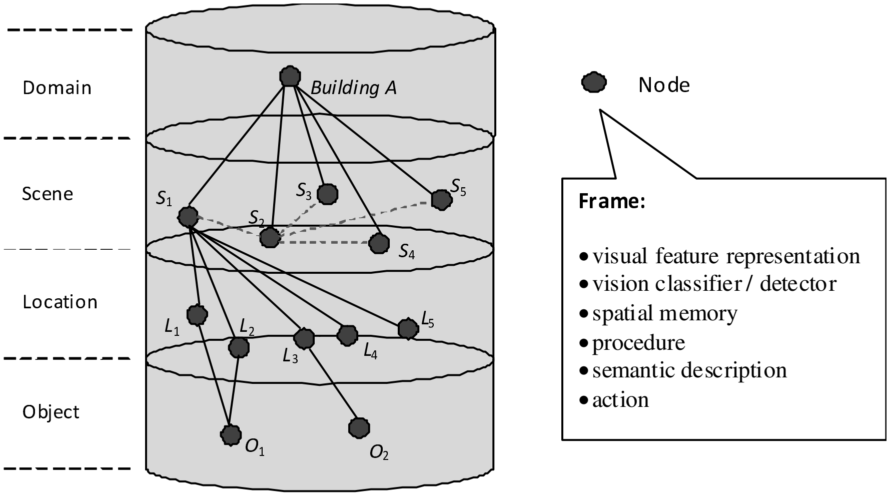


Fig. 1. Left: Long-Term Memory (LTM) model for domain knowledge representation

select a few distinctive locations for each scene, and select a few landmark objects in the scenes which are helpful for the navigation within the specific building. The hierarchical representation is edited as an XML file.

Compared with existing solutions, one can easily observe the advantages of cognitive memory representation in terms of *flexibility* and *scalability*. There is no need to input detailed floorplans or large dataset of location images. When applied to another building, the user just need to re-edit the XML file and load the trained classifiers and detectors.

2.2 Working Memory

The working memory supports the online processing for visual perception, reasoning, context-awareness, and action. The working memory is implemented according to existing understanding of working memory in human cognitive system [11,12], as shown in Figure 2. The Central Executive module performs the routine processes, attention selection, and actions for context-aware service:

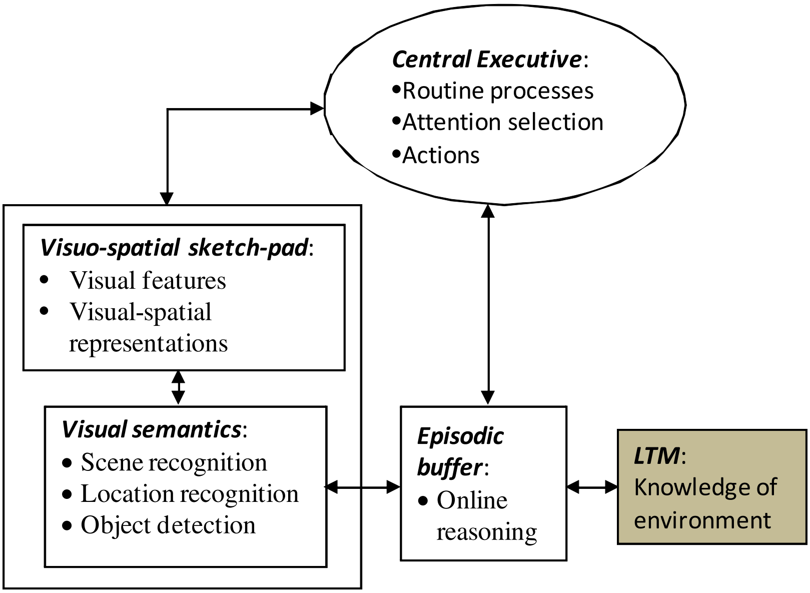


Fig. 2. Working Memory or Short-Term Memory model for online visual perception and context-aware assistant

- Routine processes: First, extract the low-level image features, such as gradients and SIFT keypoints. Then, generate the BOW (Bag-of-Words) based representation of the input image. Next, starting from the root node of LTM, perform scene recognition based on the models of each scene in the next layer.

- Attention selection: Select active node at scene level according to the recognition result and temporal reasoning on the record in episodic buffer. Then, move to the active scene node in next level of LTM to perform location recognition and select an active node of location.
- Action: according to the context-related knowledge of the active location node in LTM, detect landmark or informative objects and provide context-aware assistant according to detection output.

Data are stored in three memory modules in Working Memory, i.e. *visuo-spatial sketch-pad* for local and global image features, *visual semantics* for semantic results, and *episodic buffer* for temporal reasoning. They are connected with LTM to perform online comtextual temporal reasoning.

2.3 Efficient Visual Recognition

We employ two different appearance-based recognition schemes: image classification for scene and location recognition, and object detection for searching and locating landmark objects under the scene contextual hierarchy.

The bag-of-words (BOW) representation has been shown to be effective for image classification, but it may not be affordable for real-time tasks. In this system, a novel efficient BOW-based method proposed by us [13] is employed. Unlike conventional methods which build the bag-of-words on the raw local features *e.g.*, SIFTs, our method first maps the low-level features into an embedded manifold subspace by an improved Spectral Regression, called Expended Spectral Regression (ESR), and then generates the visual feature dictionary in the embedded feature space. We use OpenCV to generate SIFT keypoints from images. Multiple SVM classifiers are trained for image classification, where, to be efficient, linear SVM classifier is used. In our method, the dimension of word feature is about half of the original feature, and there is no need to employ SPM (Spatial Pyramid Matching) to encode spatial information. Hence, it requires much fewer memory and computational resources. The same BOW-based image representation is used for both scene and location recognition.

The Histogram-of-Oriented-Gradients (HOG) feature representation [14] is used to detect landmark objects. The integral image technique [15] is used to generate the HOG histograms, allowing the system to perform multi-scale multi-object detection very efficiently. Linear SVM classifier is again used for the efficiency. Since a single object may give rise to duplicate detections at slightly different positions and scales, clustering is used to combine these duplicates into one single detection.

2.4 Contextual Temporal Reasoning

Since the image classification is performed on a high-dimension feature space (1000 bag-of-words), the classification results would be unstable for many novel and uninformative images not included in the training set. However, temporal contextual information can help to make a correct decision. Therefore, the episodic buffer is implemented to perform temporal contextual reasoning to improve the vision-based localization.

First, a transfer time $T(l_j, l_k)$ is stored in the node in LTM, where $T(l_j, l_k)$ is the minimum time to walk from location or scene l_j to l_k. When the system is running, a record $M_{ST}[L_{ST}]$ in episodic buffer is constantly updated, where $M_{ST}[L_{ST}]$ stores the latest active scene and location nodes for about 5 minutes. The temporal contextual constraint on the current scene and location can be computed as

$$P_t(l_k) = \frac{1}{L_{ST}} \sum_{i=1}^{L_{ST}} F_k(i), \quad \text{and} \quad F_k(i) = \begin{cases} +1, & \text{if } i \geq T(l(i), l_k) \\ -1, & \text{otherwise.} \end{cases} \tag{1}$$

where $F_k(i)$ is the temporal evidence function, $l(i) = M_{ST}[L_{ST} - i]$, with $l(i) \in \{l_k\}$ and $\{l_k\}$ is the set of all scenes and locations. Combining the temporal contextual constraint and image classification results, the current scene or location can be predicted as $l_c = \arg\max_{\{l_k\}} [P_t(l_k)P_v(l_k)]$, where $P_v(l_k)$ is the confidence of vision recognition. If $\max[P_t(l_k)P_v(l_k)] \leq 0$, l_c keeps the previous label. Employing temporal contextual reasoning, one can filter out many unreasonable errors.

3 Experimental Results

A prototype of the system was built for real-world testing. It consists of a wearable web camera (Logitech HD1080p) which can be placed on the wearer's shoulder, and a Samsung Series 7 Slate tablet (1.86GHz Core i5 processor and 4GB of RAM). The system runs at about 5 frames per-second at 320×240 resolution.

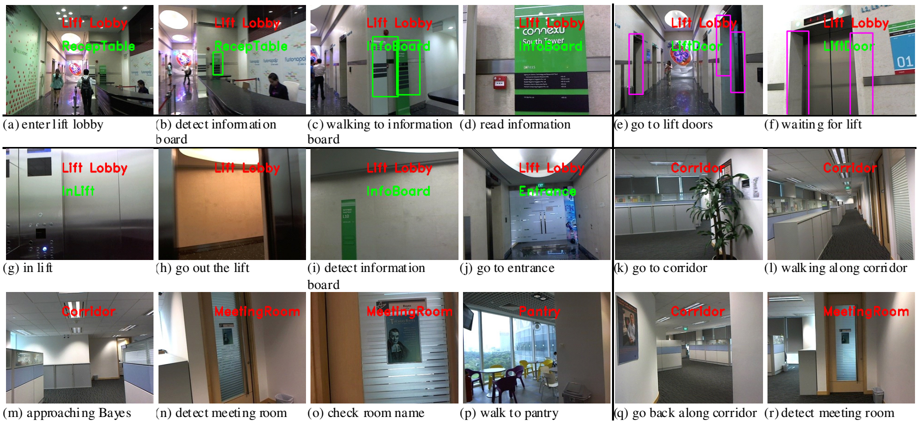
Table 1. Comparison of navigation steps with and without our system

Without our system	With our system
• enter lift lobby; • ask for the way to Bayes (**Q1**); • led to InfoBoard, find Bayes on L10; • go to lift door; • ask for which lift to take (**Q2**); • take lift to L10; • go to the entrance door; • ask for the directions to Bayes (**Q3**); • go to Bayes, find the door;	• enter lift lobby; (*system directs user to Information Board*); • read Information Board and find Bayes on L10; • go to lift door; (*system directs user to take left set of lifts*); • take lift to L10; • go to the entrance door; (*system directs user to turn right and go straight*); • go to Bayes, (*system recognizes the door*);
At least 3 enquires (or additional cognitive loads)	No enquire

The field tests were performed in South Tower, Fusionopolis, in Singapore. The building has 21 floors of office spaces, hosting several research institutes and IT companies. We classify the public and working areas in the building into five typical scene categories, *i.e.*, *Corridor*, *Cubicle*, *Lift Lobby*, *Meeting Room*, and *Pantry*. For each scene, there are a few distinctive locations which are helpful for navigation. For example, in the entrance area of *Lift Lobby* on the first floor, there is a reception table. Going further inside, there are two information boards (*InfoBoard*) where the visitor can find useful information. Further into the lift lobby, the lift doors are lined on both sides, where the lifts on one side will take passengers to the 5th to 13th floors, and the lifts on the other side will take passengers directly to the 13th floor and higher. When you go out the lift, there is an information board in the entrance to the working area. The domain knowledge is organized as in Figure 1, where *Building A* is '*South Tower, Fusionopolis*', S_1='*LiftLobby*', S_2='*Corridor*', S_3='*Cubicle*', S_4='*MeetingRoom*', and S_5='*Pantry*'. In the Location layer, L_1='*RecepTable*', L_2='*InfoBoard*', L_3='*LiftDoors*', L_4='*InLift*', and L_5='*Entrance*'. The landmark objects are information board (O_1) and lift door (O_2).

To evaluate the effectiveness of our system for indoor navigation, we tested the system in the scenario of helping a visitor to reach a meeting room, *Bayes*, on the 10th level (L10). This event is selected since the room is located in a corner and its door is not visible when you are approaching it along the long corridor until you are in front of the door, as illustrated in Figure 3 (l)-(n) where Bayes is located at the furthest end in (l). The scenarios of a visitor with and without our system would be described as those listed in Table 1.

When a visitor enters the building the first time, he may enquire which floor the meeting room *Bayes* is (**Q1**). Then, when he approaches the lifts, he may ask for which lift to take (**Q2**). As he arrives at the 10th floor and enters the working area, he may want to find a person to ask for directions to Bayes (**Q3**). If he is wearing our system, when he just enters the lift lobby, the system recognizes that the scene is *LiftLobby* and the location is in front of *RecepTable*, and then it can detect the information board, as shown in Figure 3 (b). At this time, it provides a context-aware prompt to the user verbally to go to the information board. At the information board the user will find out that Bayes is on the 10th floor (shown in Figure 3 (d)). When he goes further into the lift lobby, the system detects the lift doors (as shown in Figure 3 (e)) and prompts the user to take one of the lifts on his left to go to the 10th floor. When he gets out of the lift, the system would detect the information board (Figure 3 (i)) and prompts the user to check if he is on the 10th floor. When the user goes through the entrance and walks into the corridor (Figure 3 (j) and (k)), the system recognizes the location and prompts the user to turn right and go straight along the corridor for about 30m to Bayes. When the user is approaching Bayes, the system can detect the door of meeting room (Figure 3 (n)) and prompt the user to check if it is Bayes from the label on the door (Figure 3 (o)). If the user misses the correct room and goes further along the corridor, he will arrive at the pantry (Figure 3 (p)). The system recognizes the scene of pantry and prompts the user that he might have passed Bayes and he should turn back and go along the corridor for about a few meters to find Bayes. Once he faces the door of Bayes, the system recognizes the door of the meeting room and prompts the user to check if it is indeed Bayes, as shown in Figure 3 (r).



(a) enter lift lobby (b) detect information board (c) walking to information board (d) read information (e) go to lift doors (f) waiting for lift

(g) in lift (h) go out the lift (i) detect information board (j) go to entrance (k) go to corridor (l) walking along corridor

(m) approaching Bayes (n) detect meeting room (o) check room name (p) walk to pantry (q) go back along corridor (r) detect meeting room

Fig. 3. Typical navigation sequence for a visitor headed to meeting room *Bayes* on the 10th floor. In each example image, the red text superimposed on the image indicates the recognized scene, the green text indicates the recognized location, the green box indicates the detected information board, and the pink box indicates detected lift doors.

Three people tested the system on the scenario. In all the tests, the system performed well for context-awareness and successfully provided the online prompts for navigation as indicated in Table 1. To evaluate the effectiveness of cognitive approach, we also compared with direct image classification on errors of level-1 processing, *i.e.* scene recognition. The errors of image classification are 886 frames in three tests (total 6742 frames), while the errors of our cognitive approach have reduced to 154 frames. The performance of direct vision recognition is not stable in dynamic scenes, however, employing cognitive memory approach, the errors are reduced to less than 1/5 of direct vision method.

4 Conclusions

In this paper, we have developed a novel wearable cognitive vision system which mimic human's cognitive abilities to perform indoor navigation as a virtual companion. We propose a hierarchical knowledge structure for long-term memory which combines general concepts and specialized spatial structures of an indoor environment. We develop a working memory after human's working memory in cognitive system for real-time visual perception, which integrates scene recognition, location recognition, and multi-object/multi-scale detections. The computational cognitive system mimics the capabilities of human cognitive system to recognize a user's environment efficiently and accurately, so as to provide context-aware assistance for navigation. Wearing our system, it seems that the user is walking with a virtual companion who knows the environment well. In the near future, we want to enhance the system's cognitive capabilities for more complex tasks in human daily lives.

References

1. Sanpechuda, T., Kovavisaruch, L.: A Review of RFID Localization: Applications and Techniques. In: Proceedings of ECTI-CON, vol. 2, pp. 769–772 (2008)
2. Kashiwabara, T., Osawa, H., Shinozawa, K., Imai, M.: TEROOS: A Wearable Avatar to Enhance Joint Activities. In: Proceedings of CHI, pp. 2001–2004 (2012)
3. Mukawa, M., Lim, J.-H.: A Review of Cognitive Architectures for Visual Memory. In: Proceedings of BICA, pp. 233–238 (2012)
4. Haverinen, J., Kemppainen, A.: A Geomagnetic Field based Positioning Technique for Underground Mines. In: Proceedings of ROSE, pp. 7–12 (2011)
5. Filliat, D., Meyer, J.A.: Map-based Navigation in Mobile Robots:-I. A Review of Localization Strategies. Cognitive Systems Research 4, 243–282 (2003)
6. Meyer, J.A., Filliat, D.: Map-based Navigation in Mobile Robots:-II. A Review of Map-Learning and Path-Planning Strategies. Cognitive Systems Research 4, 283–317 (2003)
7. Pronobis, A., Caputo, B., Jensfelt, P., Christensen, H.I.: A Realistic Benchmark for Visual Indoor Place Recognition. Robotics and Autonomous Systems 58, 81–96 (2010)
8. Chong, H.-Q., Tan, A.-H., Ng, G.-W.: Integrated Cognitive Architectures: A Survey. Artif. Intell. Rev. 28, 103–130 (2007)
9. Bauchhage, C., Wachsmuth, S., Hanheide, M., Wrede, S., Sagerer, G., Heidemann, G., Ritter, H.: The Visual Active Memory Perspective on Integrated Recognition Systems. Image and Vision Computing 26, 5–14 (2008)
10. Stewart, T.C., Choo, F.-X., Eliasmith, C.: Spaun: A Perception-Cognition-Action Model Using Spiking Neurons. In: Proc. 35th Annual Conference of the Cognitive Science Society, pp. 1018–1023 (2012)
11. Baddeley, A.: Working Memory: Theories, Models, and Controversies. Annu. Rev. Psychol. 63, 1–29 (2012)
12. Schneider, W.X.: Visual-Spatial Working Memory, Attention, and Scene Representation: A Neuro-Cognitive Theory. Psychological Research 62, 220–236 (1999)
13. Li, L., Goh, W., Lim, J.-H., Pan, S.J.: Extended Spectral Regression for Efficient Scene Recognition. Pattern Recognition (submitted)
14. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Proceedings of CVPR, vol. 1, pp. 886–893 (2005)
15. Viola, P., Jones, M.: Rapid Object Detection Using a Boosted Cascade of Simple Features. In: Proceedings of CVPR, vol. 1, pp. 511–518 (2001)

Single-Image-Based Rain and Snow Removal Using Multi-guided Filter

Xianhui Zheng[1], Yinghao Liao[1,*], Wei Guo[2], Xueyang Fu[2], and Xinghao Ding[2]

[1] Department of Electronic Engineering, Xiamen University, Xiamen, China
[2] Department of Communication Engineering, Xiamen University, Xiamen, China
{yhliao,dxh}@xmu.edu.cn

Abstract. In this paper, we propose a new rain and snow removal method through using low frequency part of a single image. It is based on a key difference between clear background edges and rain streaks or snowflakes, low frequency part can obviously distinguish the different properties of them. Low frequency part is the non-rain or non-snow component. We modify it as a guidance image, the high frequency part as input image of guided filter, so we get a non-rain or non-snow component of high frequency part and add the low frequency part is the restored image. We further make it more clear based on the properties of clear background edges. Our results show that it has good performance in rain removal and snow removal.

Keywords: rain removal, snow removal, low frequency.

1 Introduction

A photo taken in the rainy day or snowy day is covered with bright streaks (Figure 1). The streaks not only cause a bad human vision, but also significantly degrade effectiveness of any computer vision algorithm, such as object recognition, tracking, retrieving and so on.

Removal of rain or snow has been paid much attention, especially rain removal. Gary and Nayar suggested a correlation model capturing the dynamics of rain and a physics-based motion blur model explaining the photometry of rain [1]. Then they proposed how to modify camera parameters to remove the effects of rain [2]. Zhang proposed a detection method combining temporal and chromatic properties of rain [5]. Barnum thought rain or snow streaks are formulated by a blurred Gaussian model, and rain or snow is detected base on the statistical information in frequency space with different frames. Then rain or snow can be removed or increased [6] [7]. Fu et al proposed a rain removal method via image decomposition, the rain component of single image could be removed via performing dictionary learning and sparse coding [8]. Jing xu proposed a method using guided filter to remove rain streaks or snow streaks[10], and then improved the performance by refining the guidance image [11].

[*] Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 258–265, 2013.

In this work, a novel method is proposed based on the difference between clear background edges and rain streaks or snow streaks. The method mainly uses the guided filter to remove rain streaks or snowflakes. Section 2 suggests that the rain streaks or snowflakes are different from other textures. In Section 3 we introduce the algorithm of removing the rain streaks or snowflakes. Section 4 shows the experimental results and compares with other methods. The last part is conclusion.

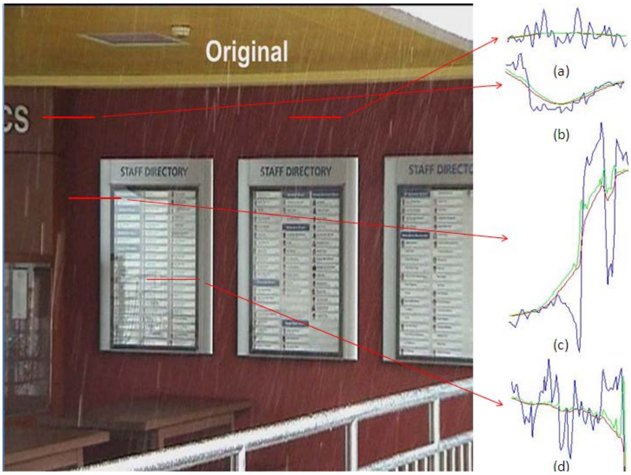


Fig. 1. Intensity profiles along the horizontal indicated by the red line. (a) Blurred rain streaks; (b) The edges with low pixel values; (c) The edges of different adjacency pixel values; (d) The clear background edges like rain. Blue line is pixel values of input image in a partial row to the same scale. The red line is the corresponding low frequency part via guided filter. The green line is the edge enhancement of red line.

2 The Difference between Clear Background Edges and Rain or Snow Streaks

Due to the size and the speed of raindrop or snowflake, they are imaged in form of bright and blurry streaks. The streaks are higher than adjacent pixel values and they will disappear in the low frequency part such as Figure 1(a). Some background edges are lower than adjacent pixel values like Figure 1(b) and they can't be rain or snow streaks. But the values will become higher through using guided filter. Other edges like Figure 1(c), some pixels near to the edges are higher, the others are lower, these edges are retained in low frequency, but become a little smooth, can be recovered through edge enhancement. There exist some textures like rain or snow streaks, which are also higher than adjacent pixel values, but they are clearer than rain or snow streaks, as shown in Figure 1(d). After transforming to low frequency part by appropriate parameters, some of them are likely to become the little textures. And we also enhance them to close to original textures.

2.1 Image Model of Blurred Rain Streaks or Snow Streaks

The commonly used mathematical model of rainy or snowy image, that is, input image can be decomposed into two components: a clean background image and a rain or snow component.

$$I_{in} = I_b + I_r \tag{1}$$

For obviously seeing the different textures of rain or snow and background, firstly, an input image is decomposed into low frequency part and high-frequency part by using guided image filter. The low frequency part is non-rain or non-snow component. All the rain and snow streaks are in the high-frequency part, which also has non-rain or non-snow textures. The opinion above is the same as [8] [12] [13]. So formula (1) can be changed to (2): I_{bl} is the low-frequency part of background. I_{bh} is the high-frequency part of background. I_{rh} denotes the rain or snow in the high-frequency part. I_{in}^{guide} means the transformation of I_{in} by using guided filter.

$$I_{in}{=}I_{bl} + I_{bh} + I_{rh} = I_{in}^{guide}{+}I_{bh} + I_r \tag{2}$$

and (2) can be simplified as (3):

$$I_{in} = I_{LF} + I_{HF} \tag{3}$$

The transformation makes the textures change, which can be showed by the red line in Figure 1. From the Figure 1(a) and Figure 3(b), we see the low frequency part is non-rain part, and contains the edges of the background. So we get a non-rain or non-snow guidance image. By using it we can remove rain and snow. The experiment results prove our idea.

2.2 Guided Filter

Guided filter is an edge-preserving smoothing filter, and has good behavior near the edges [9]. Guidance image can be itself or another reference image. Besides, "the guided filter has a fast and non approximate linear-time algorithm, whose computational complexity is independent of the filtering kernel size". So we choose it to realize our idea. Guided filter formulates output image I_{in}^{guide} is a linear of guidance image I as followed:

$$I_{in}^{guide} = \overline{a_k} I_i + \overline{b_k} \tag{4}$$

where a_k and a_k are defined as:

$$a_k = ((\sum\nolimits_{i\in\omega_k} I_i p_i)/|\omega| - \mu_k \overline{p_k}))/(\sigma_k^2 + \varepsilon) \tag{5}$$

$$b_k = \overline{p_k} - a_k \mu_k \tag{6}$$

3 Rain and Snow Removal

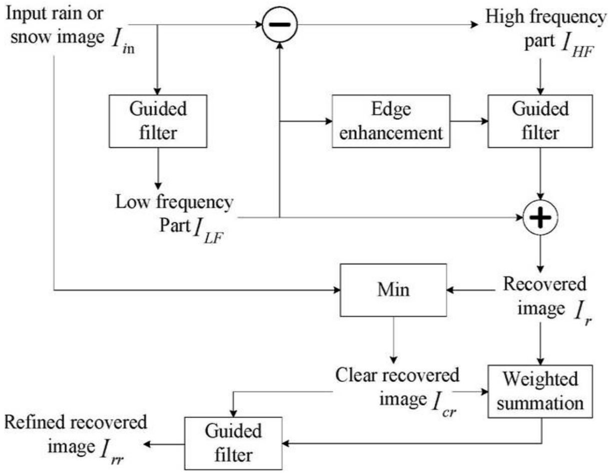

Input rain or snow image I_{in}
High frequency part I_{HF}
Guided filter
Edge enhancement
Guided filter
Low frequency Part I_{LF}
Min
Recovered image I_r
Clear recovered image I_{cr}
Weighted summation
Refined recovered image I_{rr}
Guided filter

Fig. 2. Block diagram of the proposed removal method

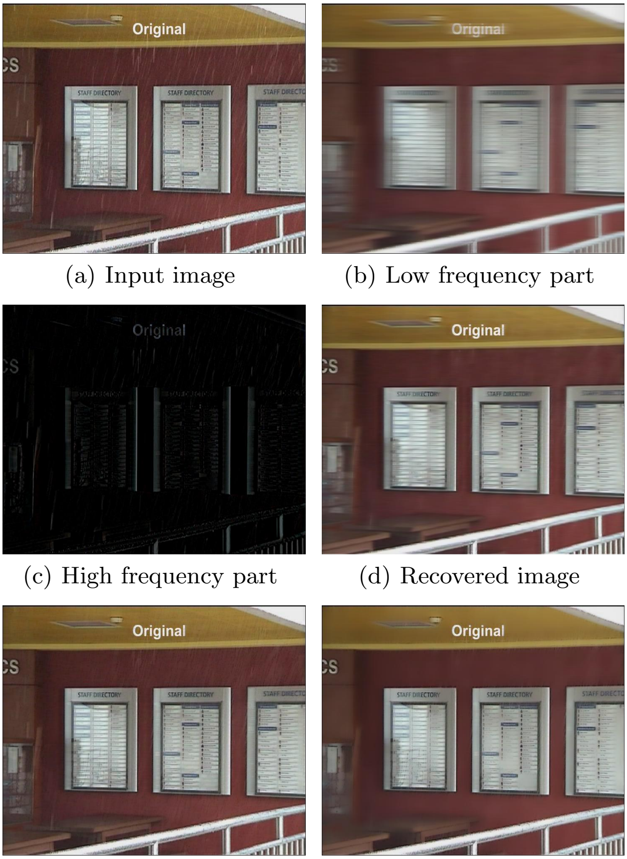
Block diagram of the proposed removal method is shown in Figure 2. It explicitly describes the framework of our method. First, input image is decomposed into low frequency part and high-frequency part by using guided filter. We introduce the low frequency part is not rain or snow part before. But due to the effect of guided filter, the edges of image become a little smooth. In order to make the existed edges more close to the edges of input image, we use edge enhancement to realize this process as follow expression:

$$I_{LF}^{*} = I_{LF} + \omega \cdot \nabla I_{LF} \tag{7}$$

where ∇I_{LF} is the gradient of I_{LF} and $\omega = 0.1$ in the paper. This enhancement is showed by the green line in Figure 1. From the Figure 1, the enhanced edges are more close to the edges of background, and all the enhanced edges are still background textures. So we get the more refined guidance image.

We don't use input image but the high-frequency part as the input image of guided filter. Since there is a big difference between the pixel values of the low frequency image and the original input image, the restored image is easy to have unsmooth flakes. The high-frequency part is more close to low frequency part, which will get better performance. After using guided filter, high-frequency part remains the non-rain or non-snow component. Through adding the low frequency part, we can get a rough recovered image.

On one hand, because we don't completely recover the edges like Fig 1(b) and just make low value pixels edges become higher, recovered image is blurred as shown in Figure 3(d). But the edges can't be rain or snow streaks, we don't



(a) Input image (b) Low frequency part

(c) High frequency part (d) Recovered image

(e) Clear recovered image (f) Refined recovered image

Fig. 3. Intermediate results in proposed method

want to change the low pixels value edges. On the other hand, using guided filter also makes recovered image blurred (such as background near to rain streaks in Figure 3(d)). So we change it to make recovered image clear as follow:

$$\mathrm{I}_{cr} = \min(I_r, I_{in}) \tag{8}$$

Due to the effect of guided filter, the values of removing rain or snow part are a little higher than nearby pixel values as shown in Figure 3(e). It is not good for our visual. So we take a weighted summation of I_r and I_{cr} to get the refined guidance image (equation 9), and then use guided filter once again to get the final result.

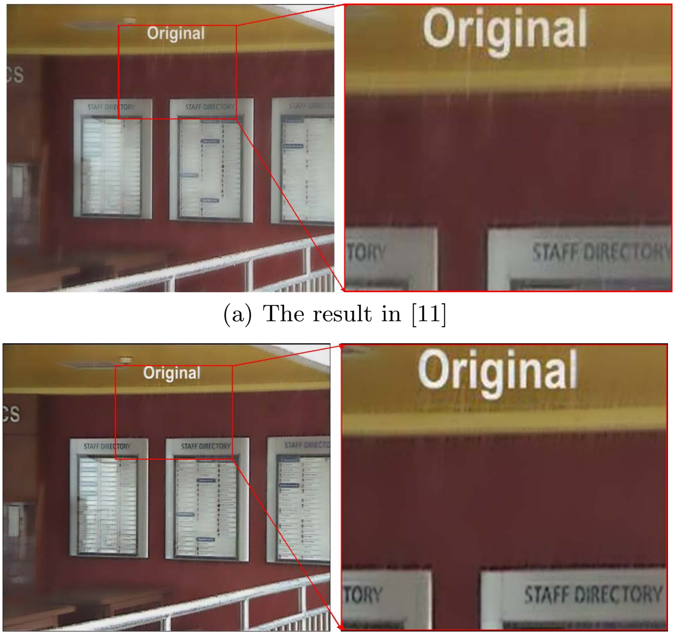
$$\mathrm{I}_{ref} = \beta I_{cr} + (1 - \beta) I_r \tag{9}$$

where $\beta = 0.8$ in rain removal and $\beta = 0.5$ in snow removal in this paper.

4 Experimental Results

Figure 3 shows the removal procedure and intermediate results proposed by this paper. Figure 3(b) is low frequency part of input image using guided filter in horizon(because the direction of rain streaks can't be in horizon). It indeed doesn't contain rain information. Figure 3(c) is high frequency part of input image. From the result, we can see that high frequency part has the information of background textures and rain streaks. And the rough recovered image Figure 3(d) is not rain but blurred because of guided filter. With minimize the input image and recovered image we get the clear recovered image Figure 3(e). It is clear, but has some flakes. We balance it with rough recovered image to get final result Figure 3(f), and it does not have flakes and very clear.

Figure 4 shows the best result of [11] and our result. Visually it is obvious that our result is better than the result of the method [11]. Our result is clearer and more effective in rain removal (e.g., the zoom-in region of red box). By the way, our method and method [11] both use guided filter, and guided filter is $O(N)$ time algorithm.

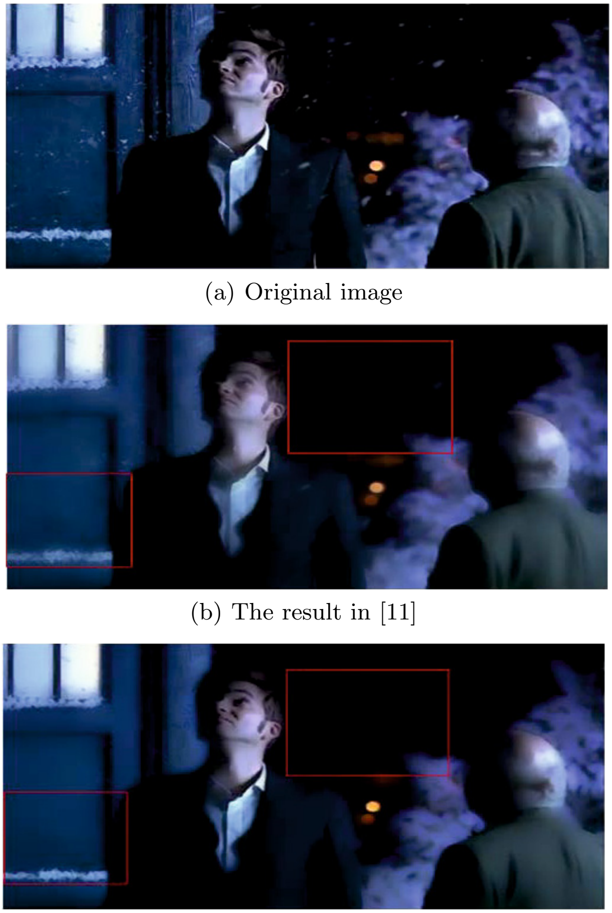


(a) The result in [11]

(b) Proposed method

Fig. 4. Comparison of rain removal results

Figure 5 shows a comparison between removal result of snow obtained by [11] and the result of our algorithm. Clearly, our method achieves more accurate removal of the snow. And our restored image is more clear (e.g., the background in the red box in Figure 5).



(a) Original image

(b) The result in [11]

(c) Proposed methods

Fig. 5. The removal results of snow

5 Conclusion

In this paper, we propose a new method for rain and snow removal of a single image. Through analysing the difference between clear background edges and rain or snow streaks, low frequency part can express the different characteristics of them, and then a rain and snow removal method base on low frequency is proposed. The removal part is mainly made up of guided filter. The results show that our method is effective and efficient in rain removal and snow removal. Our method and method [11] both use guided filter to remove rain and snow streaks, but our method has better performance.

Acknowledgments. The project is supported by the National Natural Science Foundation of China (No.30900328, 61172179,61103121), the Natural Science

Foundation of Fujian Province of China(NO: 2012J05160), The National Key Technology R&D Program(2012BAI07B06,the Fundamental Research Funds for the Central Universities(No.: 2011121051,2013121023).

References

1. Garg, K., Nayar, S.K.: Detection and removal of rain from videos. In: Proc. CVPR, vol. 1, pp. 528–535 (2004)
2. Garg, K., Nayar, S.K.: When does a camera see rain? In: Proc. CVPR, vol. 2, pp. 1067–1074 (2005)
3. Garg, K., Nayar, S.K.: Photorealistic rendering of rain streaks. Proc. SIGGRAPH 25(3), 996–1002 (2006)
4. Garg, K., Nayar, S.K.: Vision and Rain. Proc. IJCV, 3–27 (2007)
5. Zhang, X., et al.: Rain removal in video by combining temporal and chromatic properties. In: Proc. ICME, pp. 461–464 (2006)
6. Barnum, P., Kanade, T., Narasimhan, S.: Spatio-Temporal Frequency Analysis for Removing Rain and Snow from Videos. In: Proc. ICCV, pp. 1–8 (2007)
7. Barnum, P., Narasimhan, S., Kanade, T.: Analysis of rain and snow in frequency space. Proc. IJCV 86(2-3), 256–274 (2010)
8. Fu, Y.H., Kang, L.W., Lin, C.W., Hsu, C.: Single-frame-based rain removal via image decomposition. In: IEEE Int. Conf. Acoustics, Speech & Signal Processing, Prague, Czech Republic, pp. 1453–1456 (2011)
9. He, K., Sun, J., Tang, X.: Guided Image Filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 1–14. Springer, Heidelberg (2010)
10. Xu, J., Zhao, W., Liu, P., Tang, X.: Removing rain and snow in a single image using guided filter. In: Proc. CSAE, vol. 2, pp. 304–307 (2012)
11. Xu, J., Zhao, W., Liu, P., Tang, X.: An Improved Guidance Image Based Method to Remove Rain and Snow in a Single Image. CCSE Computer and Information Science Journal 5(3), 49–55 (2012)
12. Kang, L.W., Lin, C.W., Fu, Y.H.: Automatic Single-Image-Based Rain Streaks Removal via Image Decomposition. IEEE Transactions on Image Processing 21(4), 1742–1755 (2012)
13. Chen, D.Y., Chen, C.C., Kang, L.W.: Visual depth guided image rain streaks removal via sparse coding. In: Proc. ISPACS, pp. 151–156 (2012)

Image Denoising Based on Overcomplete Topographic Sparse Coding

Haohua Zhao[1], Jun Luo[2], Zhiheng Huang[1], Takefumi Nagumo[2],
Jun Murayama[2], and Liqing Zhang[1,⋆]

[1] MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems,
Dep. of Computer Science & Engineering, Shanghai Jiao Tong Univ., Shanghai, China
haoh.zhao@sjtu.edu.cn, sylvon.wong@gmail.com, zhang-lq@cs.sjtu.edu.cn
[2] SONY Corporation, Tokyo, Japan
{Jun.Luo,Takefumi.Nagumo,Jun.Murayama}@jp.sony.com

Abstract. This paper presents a novel image denoising framework using overcomplete topographic model. To adapt to the statistics of natural images, we impose sparseness constraints on the denoising model. Based on the overcomplete topographic model, our denoising system improves over previous work on the following aspects: multi-category based sparse coding, adaptive learning, local normalization, and shrinkage function. A large number of simulations have been performed to show the performance of the modified model, demonstrating that the proposed model achieves better denoising performance.

Keywords: overcomplete, sparse coding, topograph, image denoising, multi-category, adative learning, shrinkage.

1 Introduction

Human being perceives environments mainly through vision channel. Human vision system is of elegant structures to process vision information efficiently. Sparse representation is one of the strategies used in the primary vision cortex. How to use this biological plausible model to enhance image quality is one of the fundamental problems in image processing.

To enhance image quality, one of the most widely used technique is denoising, especially for natural images. Natural images often have the features that biological system can easily process for the long period of evolution.

We can use the following equation to model the noising process,

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{1}$$

where $\mathbf{X}$ is the ground truth, i.e. original image, $\mathbf{N}$ is the unknown noise, $\mathbf{Y}$ is the observed noisy image. $\mathbf{N}$ may assume certain distribution. In this paper, we assume that $\mathbf{N} \sim N(0, \sigma^2\mathbf{I})$, i.e. a Additive White Gaussian Noise (AWGN), for simplicity, where σ^2, is the known covariance of the noise.

[⋆] To whom all correspondence should be addressed.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 266–273, 2013.

In order to improve the denoising performance, researchers introduce serval prior of ground truth in denoising models. The most important one is sparse coding model. This model assumes that the ground truth can be factorized as

$$\mathbf{X} = \mathbf{A}\mathbf{s}, \tag{2}$$

where $\mathbf{A}$ is an overcomplete basis (also called dictionary), $\mathbf{s}$ is the corresponding sparse coefficient vector.

A lot of algorithms have been proposed in the literature that shows the superior performance of the sparse model Elad et al.[1] introduced an effective denoising algorithm with OMP (Orthogonal Matching Pursuit) [2] sparse coding and K-SVD[3] basis adaptation (We will call this denoising method 'K-SVD' in this paper). Some other models, such as non-local model[4], BM3D[5], CSR model[6], explored the relationship of different patches . Zhao et al.[7] applied the visual saliency model to weight the importance of the image patches in the image reconstruction process. There are other prior assumptions. One is the overcomplete topographic model[8], which is built according to the neural model and has a good performance on denoising. However, in these algorithms, we cannot use some information of the observed images, such as different statistical feature in different type of images.

In this paper, we will introduce some ideas to improve the denoising model, including multi-category based sparse coding, adaptive learnig, local normalization, and shrinkage function. A large number of simulations will be given to show the effectiveness of these modifications.

We will first review the Overcomplete Topographic Sparse Coding (OTSC) model briefly in section 2. Then we will introduce our denoising formulation in section 3 and the evaluation result in section 4. Finally, we will conclude the paper in section 5.

2 Overcomplete Topographic Sparse Coding (OTSC)

In this section, we will briefly review overcomplete representations of topographic sparse coding based on the hierarchical generative model for natural images[8]. The generating flow chart is shown in Fig.1. Certain characteristics of biological vision are used for guidance, such as a highly overcomplete early stage (V1), topographic organization and sparse distributed activity. This model yields complex cell-like receptive fields from natural images and can be used in image processing applications, such as noise removal.

Basis functions can be learned if we apply the prior of the basis coefficients considering hierarchical structure information in the noisy generative model. The objective function here is defined as the negative log-likelihood of observed data with topographic constraints.

The neighborhood function $h(i,j)$ define relations between two neurons in the vicinity, such relations is called the topography, which reflects the co-activities between the i-th and j-th components. Under the topographic assumption, the prior probability of basis coefficient can be expressed as:

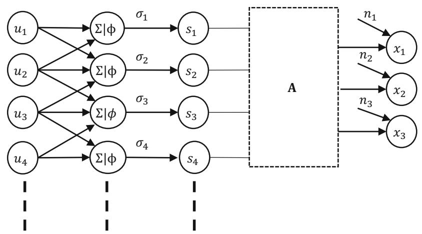


Fig. 1. Overcomplete Topographic Denoising Model

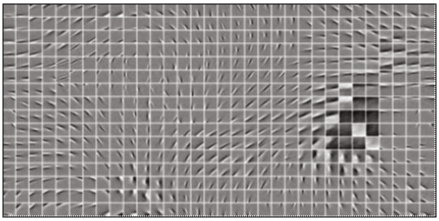


Fig. 2. Trained Overcomplete Topographic Basis Functions

$$P(\mathbf{s}) = \prod_i \exp(G(\sum_j h\,(i,j)\,s_j^2)) \tag{3}$$

where the function $G(\xi)$ has a similar role as the log-density of components in basic Independent Component Analysis (ICA). It can be of the form:

$$G(\xi) = -\alpha\sqrt{\xi + \varepsilon} + \beta, \tag{4}$$

where α is the scaling constant and β is the normalization constant.

The objective function can be obtained as follows:

$$L = \|\mathbf{x} - \mathbf{A}\mathbf{s}\|^2 - 2\sigma_n^2 \sum_i G(\sum_j h\,(i,j)\,s_j^2). \tag{5}$$

When controlling topographic representation, the second component of (5) also enforces small values of s_j^2, which results in sparsity constraint.

Optimization of this objective function can be achieved by gradient descent method in two stages: 1) To adapt basis functions to find a good coding of natural images; 2) To determine the coefficients $\mathbf{s}$ given each image $\mathbf{x}$, fixing the mixing matrix $\mathbf{A}$.

Denote $\mathbf{e} = \mathbf{x} - \mathbf{A}\mathbf{s}$ as the image residual and η as the learning rate. Given image coding vector $\mathbf{s} = (s_1, s_2, \cdots, s_n)^T$, the learning algorithm for updating basis functions is described as follow:

$$\Delta\mathbf{A} = \eta\frac{\partial L}{\partial \mathbf{A}} = -\eta\mathbf{e}\mathbf{s}^T \tag{6}$$

Once we fix the basis functions, the learning algorithm for image coding vector $\mathbf{s} = (s_1, s_2, \cdots, s_n)^T$ is given by

$$\Delta\mathbf{s} = \eta\frac{\partial L}{\partial \mathbf{s}} = \eta(2\mathbf{A}^T(\mathbf{A}\mathbf{s} - \mathbf{x}) - 2\sigma_n^2 \sum_i g(\mathbf{s}^T\mathbf{H}_i\mathbf{s})\mathbf{H}_i\mathbf{s}) \tag{7}$$

where $\mathbf{s}^T\mathbf{H}_i\mathbf{s} = \sum_{j \in G_i} h(i,j)s_j^2$.

An example of trained basis functions is shown in Fig.2.

3 Modifications of OTSC

In this section, we will introduce our modifications of OTSC model to improve the image denoising performance over the existing method in implementation.

3.1 Multi-category Based Sparse Coding

Each type of images is of its intrinsic characteristics, such characteristics can be captured by its basis functions. Therefore we need to represent images in the basis functions which match their characteristics. The idea of category division is important. With this idea, features are learned better and faster for characterizing the intrinsic features during image adaptation. Also, we can achieve higher efficiency for learning and avoid overfitting or being trapped in local maximum points during adaptation.

To verify this idea, we collected a large number of images to build a database and manually classified the images in the database into 4 categories: architecture, human being, natural scene, and plant & animal. There are 200 images in each category. We trained the basis functions for each image category by maximizing the Bayessian posterior under OTSC model. In the new approach, we denoise an image with the basis function of the corresponding category, such that the image can be more sparsely represented in the basis. Our simulation results (section 4) show that the performance of image denoising is improved with this method. This indicates a great possibility to combining with scene-recognition in camera.

3.2 Adaptive Learning

The basis function trained over a collection of images from the same category can only characterize the general features of this category. To elaborate the individual characteristics of the input image for denoising, adaptation of basis functions to it is necessary. Learning to adapt the basis is applied before denoising. It enables the basis to better describe the characteristics of the processing image. The method here is similar to training, but of different parameters, which make the procedure shorter.

To reduce cost and improve learning efficiency, we introduce a new stop criteria for adaptation. The criteria is calculated based on the variance of the noise and the difference between the sparse representation and the observation. The new adaptation process is shown in Fig. 3.

This process stops if at least one of the following conditions holds:
1. $\|\mathbf{Y} - \mathbf{A}\hat{\mathbf{s}}\|_2 < \sqrt{n}C\sigma$, or
2. The number of iterations reaches the max bound,
where $\mathbf{Y}$ is the noisy image patches, $\mathbf{A}$ is the basis functions, $\hat{\mathbf{s}}$ is the corresponding estimated coefficients, σ is the variance of the noise, n is the number of pixels of a patch, and C is a constant to control overtraining.

Our experimental simulations show that adaption correlates positively with performance, i.e. smaller C leads to better result.

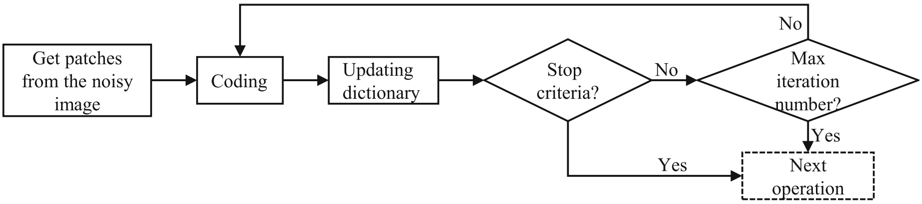


Fig. 3. Adaptive Learning

3.3 Local Normalization

The original OTSC uses (8) to prepare each patch for a better description.

$$\widetilde{x_i} = \frac{x_i - \bar{x}}{\sigma_{\text{image}}}, \tag{8}$$

where $\bar{x}$ is the mean value of all pixels in the input image. σ_{image} is the standard deviation of the whole image. This equation serves to remove the direct component of the entire image from each patch. Division by variance of the entire image adjusts each patch according to the global variance and enables sparse coding to find basis that better capture difference within the image. This eventually helps to distinguish noise from the structure of the image itself.

But there are still problems in this method. The mean values of different patches are not the same and the in-patch variance is a rather small value comparing to the mean value. A direct consequence of this is that a better part of the sparse coding has to describe the direct component of the patch. This obviously contributes negatively to our denoising system. To utilize the in-patch variance better, we apply an operation, local normalization, to the process, as shown in equation (9).

$$\hat{x}_i = \widetilde{x_i} - \bar{\bar{x}} \tag{9}$$

where $\bar{\bar{x}}$ is the mean value of pixels in patch i which has already been processed by (8). After this, sparse coding will focus on representing the in-patch variance and better prepair for denoising. Section 4 will present experiments that demonstrate the effectiveness of our approach.

3.4 Shrinkage Function

In the experiment results of the previous methods, OTSC performs well in representing image structures. But it's not good at applying a noise reduction level different from that of representing. Hyvärinen et al.[9], in a classical paper discussing image denoising by shrinkage, suggests that combining shrinkage with sparse denoising results in outstanding performance. So we try to use shrinkage function to solve the problem.

In the second component of (5), the relationship between the representation of topographic structure and the sparsity constraint cannot be changed. We put a Lasso component into the objective function (5), which turns into

$$L = \|\mathbf{x} - \mathbf{A}\mathbf{s}\|^2 - 2\sigma_n^2 \sum_i G(\sum_j h\,(i,j)\, s_j^2) + \lambda \|\mathbf{s}\|_1, \tag{10}$$

where the Lasso component is an adjustment on the sparsity constraint while keeps the topographic constraint unchanged.

Taking partial derivative of objective function (10), we get (11) and (12) for sparse coding.

$$-\frac{\partial L}{\partial \mathbf{A}} \propto (\mathbf{x} - \mathbf{A}\mathbf{s})\mathbf{s}^T \tag{11}$$

$$-\frac{\partial L}{\partial s_j} \propto \sum_i 2a_{ij} \left(x_i - \sum_j a_{ij} s_j \right) + 2\sigma_n^2 \sum_{i \in \Omega_i} g(h(i,j)s_j^2)s_j - \lambda \mathrm{sign}(s_j) \tag{12}$$

4 Simulations

To check the performance of our methods, we apply them on some of the images in our database which was also used in [7]. We choose 4 natural images from each category, 16 images in total, as test samples to simulate our method. In general case, we cannot get the noise variance. So we simulate this situation by using wavelet denoising to estimate this value. It is then treated as an input parameter to the denoising algorithm in this experiment.

We run the experiment using the original method (without adaptive learning), multi-category basis, local normalization, adaptive learning, and shrinkage function. Comparison of the result of different methods is shown on Fig. 4. A result example is shown on Fig. 5.

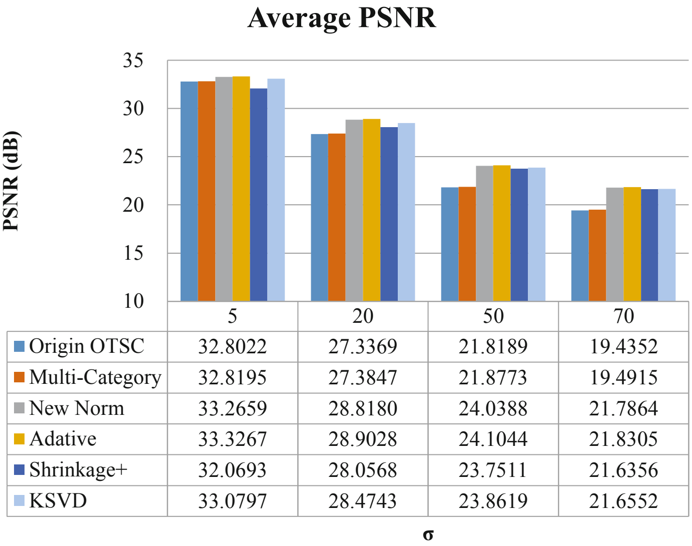
Multi-category Basis. Compared to the result using the original method, the performance improves significantly when using multi-category basis, indicating that this method yields better description of natural images.

Local Normalization. The performance is improved significantly using this method. All of the following experiments will use this method.

Adaptive Learning. We can see that the result is improved by this technique. However, since adaption is very costly, this comes at a price. We won't apply this in the following experiments.

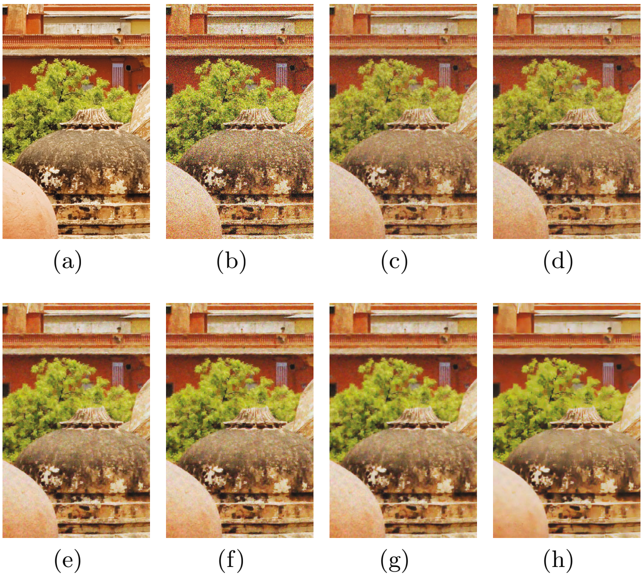
Shrinkage. We tried the case $\lambda > 0$ (Shrinkage+). From the result, we can see that more noise is removed while more detail is preserved. This indicates that a balance might be strike between topograph and sparsity when applying shrinkage function.

Finally, we apply the **K-SVD** method on the same images. Our algorithms (local norm and adaptive) have a better performance than K-SVD in the result.



	5	20	50	70
■ Origin OTSC	32.8022	27.3369	21.8189	19.4352
■ Multi-Category	32.8195	27.3847	21.8773	19.4915
■ New Norm	33.2659	28.8180	24.0388	21.7864
■ Adative	33.3267	28.9028	24.1044	21.8305
■ Shrinkage+	32.0693	28.0568	23.7511	21.6356
■ KSVD	33.0797	28.4743	23.8619	21.6552

Fig. 4. Average Peak Signal-to-Noise Ratio (PSNR) of different images. We try 4 levels of Additive Gaussian White noise. The standard deviations are 5, 20, 50, and 70.


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Example result. 5(a) Ground truth; 5(b) Noisy image (noise $\sigma = 50$); 5(c) Using basis trained by all images; 5(d) Using basis trained by multi-category; 5(e) Only local norm; 5(f) Using new adaptive learning methoed; 5(g) Increase sparse constrain; 5(h) K-SVD.

5 Conclusion

In this paper, we introduced some ideas of new framework on improving OTSC image denoising model. We proposed multi-category training basis, local normalization, adaptive learning, and shrinkage function. From the computer simulations, we found that multi-category and local normalization can greatly improve the denoising performance. So could adaptive learning, but it cost so much time. Also, we found shrinkage function can balance the relation between topographic and sparse constraints.

Acknowledgement. The work of H. Zhao and L. Zhang was in part supported by the National Natural Science Foundation of China (Grant No. 61272251, 91120305).

References

1. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing 15(12), 3736–3745 (2006)
2. Pati, Y., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 40–44 (November 1993)
3. Aharon, M., Elad, M., Bruckstein, A.: *k*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing 54(11), 4311–4322 (2006)
4. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: CVPR 2009, pp. 2272–2279 (October 2009)
5. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K., et al.: Bm3d image denoising with shape-adaptive principal component analysis. In: SPARS 2009 (2009)
6. Dong, W., Li, X., Zhang, L., Shi, G.: Sparsity-based image denoising via dictionary learning and structural clustering. In: CVPR 2011, pp. 457–464 (June 2011)
7. Zhao, H., Zhang, L.: Sparse coding image denoising based on saliency map weight. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part II. LNCS, vol. 7063, pp. 308–315. Springer, Heidelberg (2011)
8. Ma, L., Zhang, L.: Overcomplete topographic independent component analysis. Neurocomputing 71(10), 2217–2223 (2008)
9. Hyvärinen, A., Hoyer, P., Oja, E.: Image denoising by sparse code shrinkage. In: Intelligent Signal Processing. IEEE Press (1999)

Predicting Emotional States of Images Using Bayesian Multiple Kernel Learning

He Zhang, Mehmet Gönen, Zhirong Yang, and Erkki Oja

Department of Information and Computer Science
Aalto University School of Science, Espoo, Finland
{he.zhang,mehmet.gonen,zhirong.yang,erkki.oja}@aalto.fi

Abstract. Images usually convey information that can influence people's emotional states. Such affective information can be used by search engines and social networks for better understanding the user's preferences. We propose here a novel Bayesian multiple kernel learning method for predicting the emotions evoked by images. The proposed method can make use of different image features simultaneously to obtain a better prediction performance, with the advantage of automatically selecting important features. Specifically, our method has been implemented within a multilabel setup in order to capture the correlations between emotions. Due to its probabilistic nature, our method is also able to produce probabilistic outputs for measuring a distribution of emotional intensities. The experimental results on the International Affective Picture System (IAPS) dataset show that the proposed approach achieves a bette classification performance and provides a more interpretable feature selection capability than the state-of-the-art methods.

Keywords: Image emotion, low-level image features, multiview learning, multiple kernel learning, variational approximation.

1 Introduction

Affective computing [11] aims to help people communicate, understand, and respond better to affective information such as audio, image, and video in a way that takes into account the user's emotional states. Affective image classification has attracted increasing research attention in recent years, due to the rapid expansion of the digital visual libraries on the Web. In analogy to the concept of "semantic gap" that implies the limitations of image recognition techniques, the "affective gap" can be defined as "the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal" [5].

The previous research (e.g., [9,8,13]) has focused on designing features that are specific to image affect detection, after which a general-purpose classifier such as SVM [3] is used to project an image to a certain emotional category. However, the most suitable feature representation or subset related to people's emotions is not

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 274–282, 2013.

known a priori, and feature selection has to be done first for a better prediction performance in final predictions, which increases the computational complexity. Besides, an image often evokes mixed feelings in people rather than a single one, and the ground-truth labels or emotions usually conceptually correlate with each other in the affective space. In such cases, it makes more sense to assign an image several emotional labels than a single one.

In this paper, we propose a novel Bayesian Multiple Kernel Learning (MKL) method for affective image classification using low-level color, shape and texture image features. An image can be represented by different feature representations or views. MKL combines kernels calculated on different views to obtain a better prediction performance than single-view learning methods (see [4] for a recent survey). Thanks to the MKL framework, our method can learn the image feature representation weights by itself without an explicit feature selection step, which makes the interpretation easy and straightforward. Our method has been implemented within a multilabel setup in order to capture the correlations between emotions. Due to its probabilistic nature, our method is able to produce probabilistic outputs to reflect a distribution of emotional intensities for an image. The experimental results on the `International Affective Picture System (IAPS)` dataset show that the proposed Bayesian MKL approach outperforms the state-of-the-art methods in terms of classification performance, feature selection, and result interpretation.

Section 2 introduces the image features used in this paper. Section 3 gives the mathematical details of the proposed method. In Section 4, the experimental results on affective image classification are reported. Finally, the conclusions and future work are presented in Section 5.

2 Image Features

We have used a set of ten low-level color, shape, and texture features to represent each image. The features are extracted both globally and locally. Note that the features calculated for five zones employ a tiling mask, where the image area is divided into four tiles by the two diagonals of the image, on top of which a circular center tile is overlaid [12]. Table 1 gives a summary of these features. All the features are extracted using PicSOM system [6].

Four of the features are standard MPEG-7 descriptors: Scalable Color, Dominant Color, Color Layout, and Edge Histogram. 5Zone-Color is defined as the average RGB values of all the pixels within the zone. 5Zone-Colm denotes the three central moments of HSV color distribution. Edge Fourier is calculated as the magnitude of the 16×16 FFT of Sobel edge image. 5Zone-Edgehist is the histogram of four Sobel edge directions. 5Zone-Edgecoocc is the co-occurrence matrix of four Sobel edge directions. Finally, 5Zone-Texture is defined as the histogram of relative brightness of neighboring pixels.

Table 1. The set of low-level image features used

Index	Feature	Type	Zoning	Dims.
F1	Scalable Color	Color	Global	256
F2	Dominant Color	Color	Global	6
F3	Color Layout	Color	8×8	12
F4	5Zone-Color	Color	5	15
F5	5Zone-Colm	Color	5	45
F6	Edge Histogram	Shape	4×4	80
F7	Edge Fourier	Shape	Global	128
F8	5Zone-Edgehist	Shape	5	20
F9	5Zone-Edgecoocc	Shape	5	80
F10	5Zone-Texture	Texture	5	40

3 Methods

In order to benefit from the correlation between the class labels in a multilabel learning scenario, we assume a common set of kernel weights and perform classification for all labels with these weights but using a distinct set of classification parameters for each label. This approach can also be interpreted as using a common similarity measure by sharing the kernel weights between the labels.

The notation we use throughout the manuscript is given in Table 2. The superscripts index the rows of matrices, whereas the subscripts index the columns of matrices and the entries of vectors. $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. $\mathcal{G}(\cdot; \alpha, \beta)$ denotes the gamma distribution with the shape parameter α and the scale parameter β. $\delta(\cdot)$ denotes the Kronecker delta function that returns 1 if its argument is true and 0 otherwise.

Figure 1 illustrates the proposed probabilistic model for multilabel binary classification with a graphical model. The kernel matrices $\{\mathbf{K}_1, \dots, \mathbf{K}_P\}$ are used to calculate intermediate outputs using the weight matrix $\mathbf{A}$. The intermediate outputs $\{\mathbf{G}_1, \dots, \mathbf{G}_L\}$, kernel weights $\boldsymbol{e}$, and bias parameters $\boldsymbol{b}$ are used to calculate the classification scores. Finally, the given class labels $\mathbf{Y}$ are generated from the auxiliary matrix $\mathbf{F}$, which is introduced to make the inference procedures efficient [1]. We formulated a variational approximation procedure for inference in order to have a computationally efficient algorithm.

The distributional assumptions of our proposed model are defined as

$$\begin{aligned}
\lambda_o^i &\sim \mathcal{G}(\lambda_o^i; \alpha_\lambda, \beta_\lambda) && \forall(i, o)\\
a_o^i | \lambda_o^i &\sim \mathcal{N}(a_o^i; 0, (\lambda_o^i)^{-1}) && \forall(i, o)\\
g_{o,i}^m | \boldsymbol{a}_o, \boldsymbol{k}_{m,i} &\sim \mathcal{N}(g_{o,i}^m; \boldsymbol{a}_o^\top \boldsymbol{k}_{m,i}, 1) && \forall(o, m, i)\\
\gamma_o &\sim \mathcal{G}(\gamma_o; \alpha_\gamma, \beta_\gamma) && \forall o\\
b_o | \gamma_o &\sim \mathcal{N}(b_o; 0, \gamma_o^{-1}) && \forall o\\
\omega_m &\sim \mathcal{G}(\omega_m; \alpha_\omega, \beta_\omega) && \forall m
\end{aligned}$$

$$\begin{aligned} e_m|\omega_m &\sim \mathcal{N}(e_m; 0, \omega_m^{-1}) && \forall m \\ f_i^o|b_o, \boldsymbol{e}, \boldsymbol{g}_{o,i} &\sim \mathcal{N}(f_i^o; \boldsymbol{e}^\top \boldsymbol{g}_{o,i} + b_o, 1) && \forall(o,i) \\ y_i^o|f_i^o &\sim \delta(f_i^o y_i^o > \nu) && \forall(o,i) \end{aligned}$$

where the margin parameter ν is introduced to resolve the scaling ambiguity issue and to place a low-density region between two classes, similar to the margin idea in SVMs, which is generally used for semi-supervised learning [7]. As shorthand notations, all priors in the model are denoted by $\boldsymbol{\Xi} = \{\boldsymbol{\gamma}, \boldsymbol{\Lambda}, \boldsymbol{\omega}\}$, where the remaining variables by $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{b}, \boldsymbol{e}, \mathbf{F}, \mathbf{G}_1, \ldots, \mathbf{G}_L\}$ and the hyper-parameters by $\boldsymbol{\zeta} = \{\alpha_\gamma, \beta_\gamma, \alpha_\lambda, \beta_\lambda, \alpha_\omega, \beta_\omega\}$. Dependence on $\boldsymbol{\zeta}$ is omitted for clarity through-

Table 2. List of notation

N	Number of training instances
P	Number of kernels
L	Number of output labels
$\{\mathbf{K}_1, \ldots, \mathbf{K}_P\} \in \mathbb{R}^{N \times N}$	Kernel matrices
$\mathbf{A} \in \mathbb{R}^{N \times L}$	Weight matrix
$\boldsymbol{\Lambda} \in \mathbb{R}^{N \times L}$	Priors for weight matrix
$\{\mathbf{G}_1, \ldots, \mathbf{G}_L\} \in \mathbb{R}^{P \times N}$	Intermediate outputs
$\boldsymbol{e} \in \mathbb{R}^P$	Kernel weight vector
$\boldsymbol{\omega} \in \mathbb{R}^P$	Priors for kernel weight vector
$\boldsymbol{b} \in \mathbb{R}^L$	Bias vector
$\boldsymbol{\gamma} \in \mathbb{R}^L$	Priors for bias vector
$\mathbf{F} \in \mathbb{R}^{L \times N}$	Auxiliary matrix
$\mathbf{Y} \in \{\pm 1\}^{L \times N}$	Label matrix

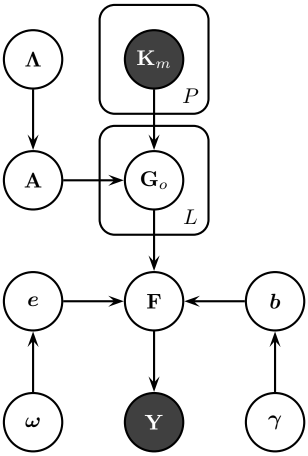


Fig. 1. Graphical model for Bayesian multilabel multiple kernel learning

out the manuscript. The variational methods use a lower bound on the marginal likelihood using an ensemble of factored posteriors to find the joint parameter distribution [2]. We can write the factorable ensemble approximation of the required posterior as

$$p(\mathbf{\Theta}, \mathbf{\Xi} | \{\mathbf{K}_m\}_{m=1}^P, \mathbf{Y}) \approx q(\mathbf{\Theta}, \mathbf{\Xi}) = \\ q(\mathbf{\Lambda})q(\mathbf{A})q(\mathbf{Z})q(\{\mathbf{G}_o\}_{o=1}^L)q(\boldsymbol{\gamma})q(\boldsymbol{\omega})q(\boldsymbol{b}, \boldsymbol{e})q(\mathbf{F})$$

and define each factor in the ensemble just like its full conditional distribution. We can bound the marginal likelihood using Jensen's inequality:

$$\log p(\mathbf{Y} | \{\mathbf{K}_m\}_{m=1}^P) \geq \\ \mathrm{E}_{q(\mathbf{\Theta}, \mathbf{\Xi})}[\log p(\mathbf{Y}, \mathbf{\Theta}, \mathbf{\Xi} | \{\mathbf{K}_m\}_{m=1}^P)] - \mathrm{E}_{q(\mathbf{\Theta}, \mathbf{\Xi})}[\log q(\mathbf{\Theta}, \mathbf{\Xi})]$$

and optimize this bound by optimizing with respect to each factor separately until convergence. The approximate posterior distribution of a specific factor $\boldsymbol{\tau}$ can be found as

$$q(\boldsymbol{\tau}) \propto \exp\big(\mathrm{E}_{q(\{\mathbf{\Theta}, \mathbf{\Xi}\} \setminus \boldsymbol{\tau})}[\log p(\mathbf{Y}, \mathbf{\Theta}, \mathbf{\Xi} | \{\mathbf{K}_m\}_{m=1}^P)]\big).$$

For our model, thanks to the conjugacy, the resulting approximate posterior distribution of each factor follows the same distribution as the corresponding factor. The exact inference details are omitted due to the space limit.

4 Experiments

In this section, we present the experimental results using our proposed Bayesian MKL method for affective image classification. We implemented our method in Matlab and took 200 variational iterations for inference with non-informative priors. We calculated the standard Gaussian kernel on each feature representation separately and picked the kernel width as $2\sqrt{D_m}$, where D_m is the dimensionality of corresponding feature representation.

4.1 Dataset and Comparison Methods

The IAPS dataset is a widely-used stimulus set in emotion-related studies. It contains altogether 1182 color images that cover contents across a large variety of semantic categories. A subset of 394 IAPS images have been grouped into 8 discrete emotional categories based on a psychophysical study [10], including Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear and Sad(ness). The ground truth label for each image was selected as the category that had majority of the votes. Both Machajdik *et al.* [9] and Lu *et al.* [8] used this subset for image emotion classification, hence we used it to compare with their results in [9,8].

4.2 Experimental Setup

We used the same training and testing procedure (80% samples for training, 20% for testing) as in [9,8]: we ran 5-fold Cross-Validation (CV) and calculated the average classification accuracy. As a baseline method, the standard SVM (with Gaussian kernel and 5-fold CV) was also implemented for comparison, where each feature was taken separately for training a single classifier.

4.3 Results

Figure 2 shows the classification results. It is clear to see that our proposed approach is the best among the three. With rather generic low-level image features, our classifier can achieve very good classification performance. Note that the compared methods [9,8] utilize complicated domain-specific features.

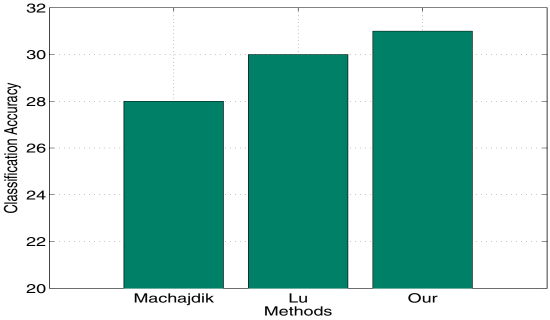


Fig. 2. The classification results of the compared methods

To further demonstrate the advantage of multiple kernel (multiview) learning over single kernel (single-view) learning, we trained and tested a single SVM classifier using each of the 10 features separately (with the same partition as MKL setup). Table 3 lists the classification accuracies. The best SVM classifier (trained with Dominant Color) can only achieve an accuracy of 0.22, which is about 9 percent lower than that of our method. And an SVM using all 10 features can give an accuracy of 0.25. This demonstrates the advantage of multiview learning over single-view learning. It also validates the strength of our proposed classifier in terms of mapping low-level image features to high-level emotional responses.

Another advantage of our MKL method is that it can select features automatically without explicit feature extraction and selection procedures. Figure 3 shows the average feature representation weights (i.e., kernel weights) in the range $[0, 1]$ based on 5-fold CV for the multiple kernel learning scenario. We clearly see that, among the ten image feature representations, Edge Histogram (F6) ranks first, followed by Scalable Color (F1), 5Zone-Colm (F5), and Edge

Table 3. The image features ranked by SVM classification accuracies

Rank	Feature	Accuracy
1	Dominant Color	0.22
2	Color Layout	0.22
3	Edge Fourier	0.22
4	5Zone-Texture	0.21
5	5Zone-Colm	0.21
6	Scalable Color	0.20
7	5Zone-Color	0.20
8	5Zone-Edgecoocc	0.20
9	5Zone-Edgehist	0.19
10	Edge Histogram	0.18

Fourier (F7) etc. This reveals that colors and edges of an image are the most informative features for emotions recognition, which is in agreement with the studies in [9] and [8]. This also shows that multiple kernel learning helps to identify the relative importances of feature representations using a common set of kernel weights.

It is worth emphasizing that an image can evoke mixed emotions instead of a single emotion. Our Bayesian classifier is capable of producing multiple probabilistic outputs simultaneously for an image, which allows us to give the image a “soft” class assignment instead of a “hard” one. This characteristic is particularly useful for detecting emotion distribution evoked by an image. Figure 4 gives some examples. One can see that the probabilistic outputs of our Bayesian classifier generally agree well with the real human votes for certain images.

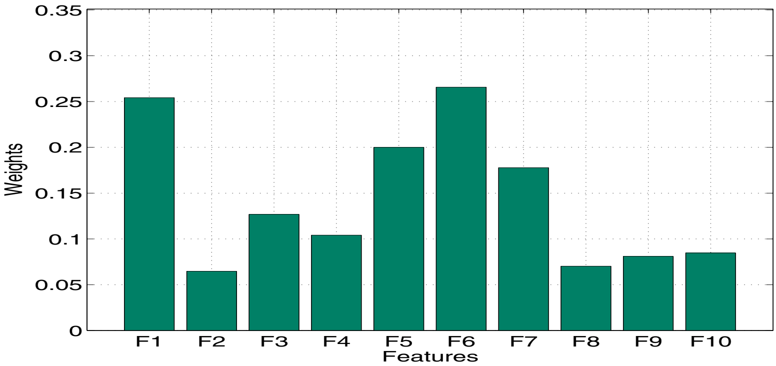
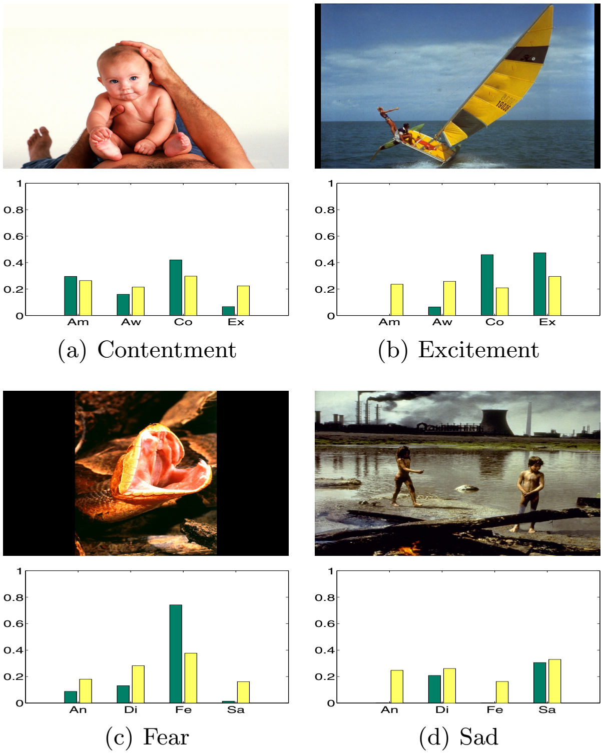


Fig. 3. The average feature representation weights over 5-fold cross-validation for the multilabel multiple kernel learning scenario



(a) Contentment

(b) Excitement

(c) Fear

(d) Sad

Fig. 4. The agreement of image emotion distribution between our predicted results (green bars) and the normalized human votes (yellow bars). The x-axis shows positive emotions ((a) & (b)): Amusement, Awe, Contentment, Excitement, and negative emotions ((c) & (d)) Anger, Disgust, Fear, Sad. The y-axis shows the agreement in the range $[0, 1]$.

5 Conclusions

In this paper, we have presented a novel Bayesian multiple kernel learning method for affective image classification with multiple outputs and feature representations. Instead of single feature (view) representation, our method adopts a kernel-based multiview learning approach for better prediction performance and interpretation, with the advantage of selecting or ranking features automatically. To capture the correlations between emotions, our method has been implemented within a multilabel setup. Due to its probabilistic nature, the proposed approach is able to produce probabilistic outputs for measuring the intensities of a distribution of emotions evoked by an image. More large-scale emotional datasets will be tested in the future. It is worth emphasizing that our method is not confined to the image recognition, but can be easily extended to other affective stimuli such as audio and video data.

Currently, only the conventional low-level image features are utilized, as our focus in this paper is not on the affective feature design. Rather, we would like to provide a new framework for better predicting people's emotional states, especially when an image evokes multiple affective feelings in people. Eventually, the development in this interdisciplinary area relies on the joint efforts from, for instance, artificial intelligence, computer vision, pattern recognition, cognitive science, psychology, and art theory.

Acknowledgements. This work has received funding from the Academy of Finland in the project Finnish Center of Excellence in Computational Inference Research (COIN). We gratefully acknowledge the *Center for the Study of Emotion & Attention* at University of Florida for providing the original IAPS image dataset.

References

1. Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88(422), 669–679 (1993)
2. Beal, M.J.: Variational Algorithms for Approximate Bayesian Inference. Ph.D. thesis, The Gatsby Computational Neuroscience Unit, University College London (2003)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
4. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2211–2268 (2011)
5. Hanjalic, A.: Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Processing Magazine 23(2), 90–100 (2006)
6. Laaksonen, J., Koskela, M., Oja, E.: PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. IEEE Transactions on Neural Networks 13(4), 841–853 (2002)
7. Lawrence, N.D., Jordan, M.I.: Semi-supervised learning via Gaussian processes. In: Advances in Neural Information Processing Systems 17, pp. 753–760 (2005)
8. Lu, X., Suryanarayan, P., Adams Jr., R.B., Li, J., Newman, M.G., Wang, J.Z.: On shape and the computability of emotions. In: Proceedings of the International Conference on Multimedia, pp. 229–238 (2012)
9. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the International Conference on Multimedia, pp. 83–92 (2010)
10. Mikels, J., Fredrickson, B., Larkin, G., Lindberg, C., Maglio, S., Reuter-Lorenz, P.: Emotional category data on images from the International Affective Picture System. Behavior Research Methods 37(4), 626–630 (2005)
11. Picard, R.: Affective Computing. MIT Press (1997)
12. Sjöberg, M., Muurinen, H., Laaksonen, J., Koskela, M.: PicSOM experiments in TRECVID 2006. In: Proceedings of the TRECVID 2006 Workshop (2006)
13. Zhang, H., Augilius, E., Honkela, T., Laaksonen, J., Gamper, H., Alene, H.: Analyzing emotional semantics of abstract art using low-level image features. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 413–423. Springer, Heidelberg (2011)

Local Linear Spectral Hashing

Kang Zhao, Dengxiang Liu, and Hongtao Lu

MOE-Microsoft Laboratory for Intelligent Computing and Intelligent
Systems Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, 200240, P.R. China
{sjtuzk,xiangzi777,htlu}@sjtu.edu.c

Abstract. Hashing for large scale image retrieval has become more and more popular because of its improvement in computational speed and storage reduction. Spectral Hashing (SH) is a very efficient unsupervised hashing method through mapping similar images to similar binary codes. However, it doesn't take the non-neighbor points into consideration, and its assumption of uniform data distribution is usually not true. In this paper, we propose a *local linear spectral hashing* framework that minimizes the average Hamming distance with a new local neighbor matrix, which can guarantee the mapping not only from neighbor images to neighbor codes, but also from non-neighbor images to non-neighbor codes. Based on the framework, we utilize three linear methods to handle the proposed problem, including orthogonal hashing, non-orthogonal hashing, and sequential hashing. The experiments on two huge datasets demonstrate the efficiency and accuracy of our methods.

Keywords: Image Retrieval, Hamming Distance, Spectral Hashing, Eigenvalue Decomposition.

1 Introduction

Content based image retrieval (CBIR) has recently attracted more and more attention due to tremendous growth of images on the internet and the growing user demand. For a typical CBIR system, it will return the nearest neighbors from a specific database of images with a pre-defined feature space and similarity metric, after users directly input an image. Because of the large quantities of images in the database, fast and accurate ways of finding nearest neighbors (NN) are in urge demand.

Some popular fast nearest neighbor search methods, including $k-d$ trees [1], M-trees, metric trees [2], are based on tree structures. These techniques aim to speed up nearest neighbors search and achieve promising performance in real applications. However, tree-based methods can not deal with large-scale data since they are essentially quite memory demanding. Moreover, the search performance may degenerate as the high dimensionality of data samples.

Recently, hashing based nearest neighbors search techniques [3–8] have attracted considerable attention. They can achieve fast query time and reduce storage requirement. Semantic hashing [8] learns compact binary codes to make

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 283–290, 2013.

similar items will have similar binary codewords and uses a simple feedforward network to calculate the binary code for a novel input. Locality Sensitive Hashing (LSH) [3] is one of the most popular unsupervised hashing methods which makes use of random projection matrixes to get binary codes. After that, some variants of LSH have been proposed, including LSH based on p-stable distributions [4], mahalanobis distance [9], and kernelized LSH [5]. Another popular and efficient unsupervised method is spectral hashing(SH) which was proposed in [6]. SH shows that learning a good compact code is equivalent to a particular form of graph partition. Therefore, with suitable relaxation, SH can be solved by the state-of-the-art efficient eigen-decomposition solvers.

Traditional unsupervised hashing methods only focus on mapping similar samples to similar codes, they don't take non-neighbors data points into account. In this paper, we propose a *local linear spectral hashing* (LLSH) framework that can map neighbor items to neighbor codes, meanwhile non-neighbor items to non-neighbor codes, which is guaranteed by formulating the hashing problem as minimizing the average Hamming distance and introducing negative values into the new local neighbor matrix. With some relaxations, we find the final formulation can be solved as a simple eigenvalue problem. Moreover, in order to get better codes, we adopt non-orthogonal and sequential methods to relax the orthogonality constraints.

The rest of this paper is organized as follows. Section 2 reviews spectral hashing. We propose our local linear spectral hashing in Section 3. The experimental evaluations are presented in Section 4. Followed with conclusion in Section 5.

2 Spectral Hashing

As previously mentioned, SH seeks the codes of data points so that it can preserve sample similarity in the hamming space. Furthermore, it requires each bit of the codes to be balanced and uncorrelated. The problem of SH can be expressed as following:

$$\begin{aligned} &\min \sum_{i,j} sim(\mathbf{x}_i, \mathbf{x}_j) \|\mathbf{Y}_i - \mathbf{Y}_j\|^2 \\ \text{subject to : } &\mathbf{Y} \in \{-1, 1\}^{n \times r}, \mathbf{1}^T\mathbf{Y} = 0, \mathbf{Y}^T\mathbf{Y} = n\mathbf{I}_{r \times r} \end{aligned}$$

where $\mathbf{Y}$ is the r-bit codes of n points, $\mathbf{Y}_i$ is the i^{th} row of $\mathbf{Y}$.

For $r = 1$, solving this problem is equal to balanced graph partitioning, and it's a NP hard problem. The condition of r-bit will make it more difficult. After spectral graph analysis [10] and the assumption of uniform data distribution, a closed solution [6] is proposed to overcome the *out of sample extension* problem. SH has been certified to be a very efficient unsupervised hashing method. Nonetheless, the PCA projections selected by SH are likely replicated when the dimensions of data are very high. Besides, for most real-world data, they hardly obey the uniform distribution.

3 Local Linear Spectral Hashing

3.1 Formulation

The purpose of this paper is to seek binary codes that satisfy (1) neighbors in the feature space are mapped to similar codes; (2) non-neighbors are mapped to codes that are far away. Suppose we have a data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$. We assume $\mathbf{X}$ is zero mean, i.e., $\sum_{i=1}^{n} \mathbf{x}_i = 0$. Let $\mathbf{y}_i (\in \mathbb{R}^k)$ be the k-bit binary code of data $\mathbf{x}_i$ and $\mathbf{A}_{n\times n}$ be the neighbor matrix, then we obtain the minimization problem:

$$\begin{aligned} \min : \sum_{i,j} A_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \\ \text{subject to} : \mathbf{y}_i \in \{-1, 1\}^k \\ \sum_i \mathbf{y}_i = 0 \\ \frac{1}{n}\sum_i \mathbf{y}_i \mathbf{y}_i^T = \mathbf{I} \end{aligned} \tag{1}$$

Although our problem has the same form as SH, there are several differences between our approach and SH: (1) the critical difference is the neighbor matrix $\mathbf{A}$. We calculate $\mathbf{A}$ with L_2 norm and set the $10th$ percentile distance in $\mathbf{A}$ as a threshold, which is used to judge neighbors and non-neighbors. More specifically, $-1 \le A_{ij} \le 1$. $A_{ij} > 0$, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbors; $A_{ij} < 0$, otherwise. By minimizing the average Hamming distance, the positive values in $\mathbf{A}$ will ensure neighbor samples are mapped to neighbor codes, while the negative values will ensure non-neighbor samples are mapped to non-neighbor codes. (2) $\mathbf{A}$ in SH is $n \times n$, which is intractable when n is very large. To overcome the computational bottleneck, we use a local neighbor matrix, whose size is $m \times m (m \ll n)$. (3) We try to find the linear solutions to the above problem, which are distinct from the nonlinear solutions of SH. We will show our linear methods in Section 3.2.

3.2 Linear Solution

As mentioned above, we try to obtain the solutions from the perspective of linear projection. That is, we want to learn a projection matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_k] \in \mathbb{R}^{d\times k}$. For every data point $\mathbf{x}_i$, the hash function of length k is defined by $\mathbf{y}_i = sgn(\mathbf{W}^T\mathbf{x}_i)$, and the $0-1$ code can be given by $\frac{1}{2}(\mathbf{1}+\mathbf{y}_i)$. $sgn(\cdot)$ will do element-wise calculation for a matrix or a vector.

With this notation, Eq.(1) can be written as:

$$H(\mathbf{W}) = \frac{1}{2} tr\{sgn(\mathbf{W}^T\mathbf{X}^T)(\mathbf{D} - \mathbf{A}) sgn(\mathbf{X}\mathbf{W})\} \tag{2}$$

where $\mathbf{D} = diag(\mathbf{A}\mathbf{1})$, $\mathbf{1} = [1, ..., 1]^T \in \mathbb{R}^n$.

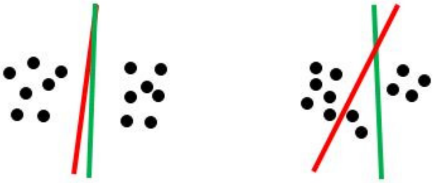


Fig. 1. An illustration of splitting the data points with hashing hyperplanes. On the left figure, the red hyperplane and the green one are equivalent. However, on the right, the green hyperplane is more reasonable.

So, we can rewrite the objective function and constraints as:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} H(\mathbf{W}) \quad (3)$$
$$\text{subject to}: \quad \mathbf{1}^T\mathbf{Y} = 0, \mathbf{Y}^T\mathbf{Y} = n\mathbf{I}_{k\times k}$$

where $\mathbf{Y} = sgn(\mathbf{XW}) \in \mathbb{R}^{n\times k}$.

3.3 Relaxing Objective Function

Inspired by Semi-Supervised Hashing(SSH) [7], we substitute signed magnitude of projection for its sign. Eq.(2) is continuous now:

$$H(\mathbf{W}) = \frac{1}{2}tr\{\mathbf{W}^T\mathbf{X}^T(\mathbf{D} - \mathbf{A})\mathbf{XW}\} \quad (4)$$

The constraints $\mathbf{1}^T\mathbf{Y} = 0$ mean each bit takes 50% probability to be 1 or -1. However, for real-world data, it is not always be accepted, as illustrated in Figure 1. The constraints will force one to select the red hyperplane instead of the green one. Hence, we will ignore the constraint.

Next, we relax the uncorrelation of bits $\mathbf{Y}^T\mathbf{Y} = n\mathbf{I}_{k\times k}$ by imposing the constraints $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{k\times k}$, which request the projection directions to be unit-norm and orthogonal to each other. Now, we have the relaxed problem:

$$\min: H(\mathbf{W}) = \frac{1}{2}tr\{\mathbf{W}^T\mathbf{MW}\} \quad (5)$$
$$\text{subject to}: \mathbf{W}^T\mathbf{W} = \mathbf{I}$$

where $\mathbf{M} = \mathbf{X}^T(\mathbf{D} - \mathbf{A})\mathbf{X}$. This is a typical eigenvalue decomposition problem, whose solutions are the eigenvectors corresponding to the k smallest eigenvalues of $\mathbf{M}$.

If we regard the hashing problem as splitting the feature space with hyperplanes, the orthogonality constraints on the projection directions mean that the hyperplanes are orthogonal to each other, which are too rigorous. The optimal hyperplane is not necessarily orthogonal to all of the rest, we empirically certificate this in Section 4. Therefore, we convert the hard orthogonality constraints into a penalty term added to the objective function:

$$H(\mathbf{W}) = \frac{1}{2}tr\{\mathbf{W}^T\mathbf{MW}\} + \rho\|\mathbf{W}^T\mathbf{W} - \mathbf{I}\|_F^2 \quad (6)$$

Algorithm 1. The modified sequential projection learning for hashing (M_SPLH)

```
Input: data $\mathbf{X}$, the original matrix $\mathbf{L}_1(\mathbf{A}–\mathbf{D})$, length of hash codes $K$
for $k = 1$ to $K$ do
    Computer the target matrix:
        $\mathbf{M}_k = \mathbf{X}^T\mathbf{L}_k\mathbf{X}$
    Extract the eigenvector $e$ of the largest eigenvalue of $\mathbf{M}_k$ and set:
        $\mathbf{w}_k = e$
    Calculate the $k^{th}$ projection:
        $\tilde{\mathbf{L}}^k = \mathbf{X}\mathbf{w}_k\mathbf{w}_k^T\mathbf{X}^T$
    Update the neighbor matrix:
        $\mathbf{L}_{k+1} = \mathbf{L}_k- T(\tilde{\mathbf{L}}^k, \mathbf{L}_k)$
end for
```

The positive coefficient ρ is used to adjust the tolerance to non-orthogonality. To get the solution, we adopt the same technique mentioned in [7].

In addition, considering LLSH-nonorth may be sensitive to coefficient ρ, we further propose a solution: LLSH-splh, which is a variant of sequential projection learning hashing [7]. It can generate robust codes by iteratively updating the neighbor matrix. Algorithm 1 will show you the structure of our modified version. The function $T(\cdot)$ means:

$$T(\tilde{\mathbf{L}}_{ij}^k, \mathbf{L}_{ij}) = \begin{cases} sgn(\tilde{\mathbf{L}}_{ij}^k) & \text{if } sgn(\tilde{\mathbf{L}}_{ij}^k \cdot \mathbf{L}_{ij}) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\mathbf{L}_{ij}$ is the element of the i^{th} row and j^{th} column in $\mathbf{L}$.

4 Experiments

We evaluate the three solutions of LLSH: LLSH-orth, LLSH-nonorth and LLSH-splh on the CIFAR dataset and STL dataset, and compare our results with SH.

4.1 CIFAR-10 Dataset

The CIFAR-10 dataset is a labeled subset of the 80 million tiny images dataset. It consists of 60000 32×32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.[1] Each image is associated with only one class label from 1 to 10. We use 50000 images as the training set, and 1000 images as the test set. For LLSH, 2000 data points are randomly sampled from training set to construct the neighbor matrix $\mathbf{A}$. We convert the original images to 512-dim GIST feature. The retrieval results are assessed with two criteria: Hamming ranking and Hash lookup as in [7]. For Hamming ranking, we record the precision on the top 500 returned samples. For Hash lookup, we evaluate the precision within hamming radius 2(including 2). Figure 2 show

[1] http://www.cs.toronto.edu/~kriz/cifar.html

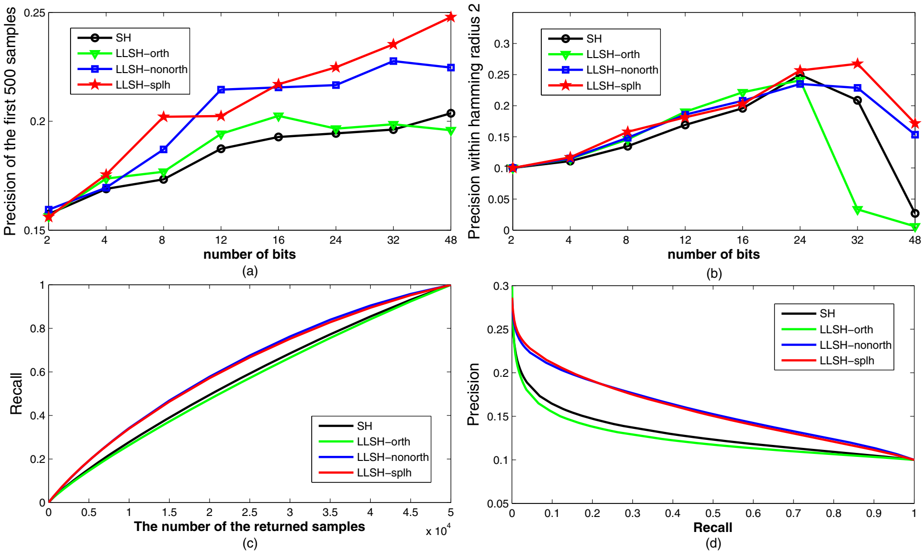


Fig. 2. The Experimental results on CIFAR10 dataset. a) Hamming ranking precision of first 500 returned samples; b) Hash lookup precision within hamming radius 2; c) The recall curves of different numbers of the returned samples using 48-bit codes; d) The precision-recall results using 48-bit codes.

the experimental results in details. We alter the hash bits from 2 to 48. From the figure, LLSH-orth and SH have the comparable performance, but LLSH-nonorth significantly performs better especially for higher bits because of the contributions of non-orthogonality, as well as LLSH-splh, which can iteratively rectify the mistakes made by the previous hash function.

4.2 STL-10 Dataset

The STL-10 dataset is an image recognition dataset for developing unsupervised learning algorithms. It has 10 classes, 5000 training images (500 images per class), 8000 test images (800 images per class) and 100000 unlabeled images.[2] In our experiments, the training set consists of 100000 images, and the test set consists of 1000 images. For LLSH, we derive the neighbor matrix $\mathbf{A}$ using a separate set of 2000 samples from the training set. Because STL10 dataset is unlabeled, we randomly pick 10000 samples from the training set to construct a pair-wise distance matrix $\mathbf{D}^*$ with L_2 norm and set the $10th$ percentile distance in $\mathbf{D}^*$ as the threshold, which is used to judge good neighbors from a query. We perform hashing coding in the 384-dim GIST feature space. Figure 3 shows the quantitative evaluation of different methods. The too strict orthogonality

[2] http://www.stanford.edu/~acoates/stl10/

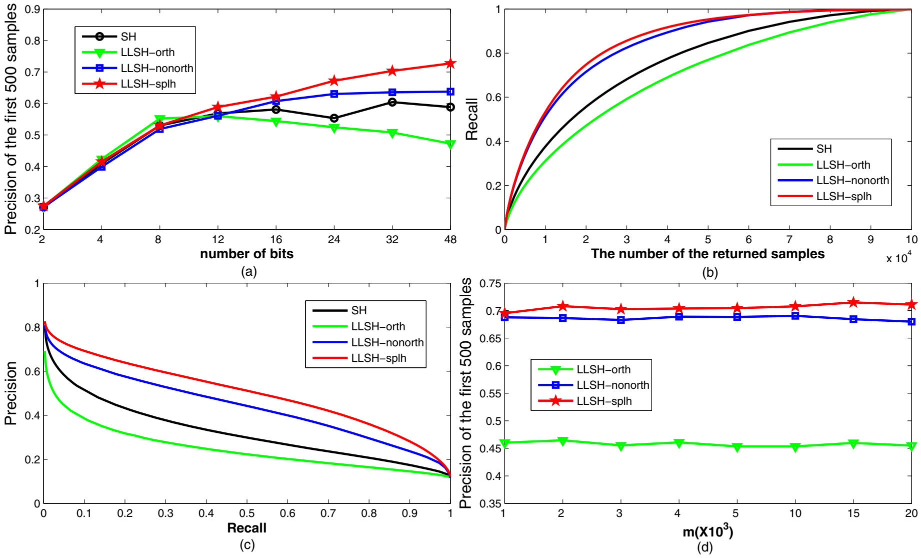


Fig. 3. The Experimental results on STL10 dataset. a) Hamming ranking precision of first 500 returned samples; b) The recall curves of different numbers of the returned samples using 48-bit codes; c) The precision-recall results using 48-bit codes; d) Hamming ranking precisions of different m on STL10 dataset using 48-bit codes.

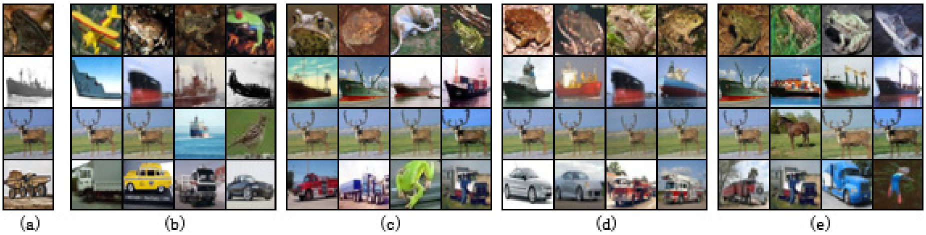


Fig. 4. The retrieval results of different methods using 48-bit codes. a) Query images; b) LLSH-orth; c) LLSH-nonorth; d) LLSH-splh; e) SH.

constraints may partly explain the bad performance of LLSH-orth. However, LLSH-nonorth and LLSH-splh both outperform SH when the number of bits is higher. And in most cases, LLSH-splh performs best.

Besides, we exhibit the hamming ranking precisions of different m in Figure 3(d). The results show that our experiments are not sensitive to m. Finally, Figure 4 shows the top 4 returned images of our methods and SH on some sample queries. It is clear that LLSH-nonorth and LLSH-splh tend to have better performance.

5 Conclusions

In this paper, we have proposed a LLSH framework to learn hash codes, which minimizes the average Hamming distance with a new local neighbor matrix so as to preserve similarity/dissimilarity in feature and hamming space. After relaxing the objective function, we can directly solve the problem by eign-decomposition. In order to improve the accuracy of bits, we further take advantage of another two techniques to attain the solutions. The experiments on two huge datasets exhibit that our method achieves promising performance.

References

1. Silpa-Anan, C., Hartley, R.: Optimised kd-trees for fast image descriptor matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
2. Uhlmann, J.K.: Satisfying general proximity/similarity queries with metric trees. Information Processing Letters 40(4), 175–179 (1991)
3. Gionis, A., Indyk, P., Motwani, R., et al.: Similarity search in high dimensions via hashing. In: Proceedings of the International Conference on Very Large Data Bases, pp. 518–529 (1999)
4. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth Annual Symposium on Computational Geometry, pp. 253–262. ACM (2004)
5. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing for scalable image search. In: IEEE 12th International Conference on Computer Vision, pp. 2130–2137 (2009)
6. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Advances in Neural Information Processing Systems, pp. 1753–1760 (2008)
7. Wang, J., Kumar, S., Chang, S.: Semi-supervised hashing for large scale search. IEEE Trans. on Pattern Analysis and Machine Intelligence 34(12), 2393–2406 (2012)
8. Salakhutdinov, R., Hinton, G.: Learning a nonlinear embedding by preserving class neighbourhood structure. In: AI and Statistics, vol. 3, p. 5 (2007)
9. Kulis, B., Jain, P., Grauman, K.: Fast similarity search for learned metrics. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(12), 2143–2157 (2009)
10. Fowlkes, C., Belongie, S., Chung, F., Malik, J.: Spectral grouping using the nystrom method. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(2), 214–225 (2004)

Latency Modulation of Border Ownership Selective Cells in V1-V2 Feed-Forward Model

Ko Sakai and Shunsuke Michii

Department of Computer Science, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
sakai@cs.tsukuba.ac.jp

Abstract. Border Ownership (BO) selective cells show complex behavior in latency, such as a shorter latency for a farther figural element. The complex behavior of latency would give crucial constraints for understanding the neuronal mechanisms underlying BO selectivity. We propose that the delays in the feed-forward mechanisms of surround modulation account for the characteristics of the latency. Specifically, combinations of delays between V1 cells, surround center and BO-selective cells underlie the latency characteristics. To examine the hypothesis, we constructed the V1-V2 feed-forward model and carried out Monte Carlo simulations with a variety of surround regions. The results indicate that physiologically realistic surround regions reproduce the puzzled behavior of latency. The results support the feed-forward mechanism of surround modulation for the neural mechanism underlying the BO selectivity.

Keywords: vision, perception, figure-ground, surround modulation, latency.

1 Introduction

Cortical representation of object in the inferiortemporal cortex gradually emerges from early representations of contour, surface, and primitive shape through early-to-intermediate-level visual areas such as V1, V2 and V4. The construction of surface appears to be crucial in the integration of the components and features. Border ownership (BO) is a fundamental element in surface construction, which defines the direction of figure along contours. Physiological studies have reported that a majority of monkey V2 cells are selective to BO [e.g., 1]. The response of BO-selective cells is modulated by stimuli that fall on to the outside of the classical receptive field (CRF). A recent study on cellular dynamics has reported that latency of BO-selective cells is modulated by the location (with respect to the CRF) and size of stimulus in a complex manner. Specifically, a shorter latency is observed for a smaller stimulus and a figural element located farther away, and a longer latency for a larger stimulus and a nearer element. This behavior of the latency would give crucial constraints for understanding the neuronal mechanisms underlying BO selectivity and figure-ground segregation. Sakai and Nishimura [e.g., 3, 6] have proposed a computational model of BO-selective cells based on surround modulation of early visual areas, which reproduced successfully a number of physiological observations. However, they focused on the

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 291–300, 2013.

spatial characteristics of the cells, and thus anatomical connections and latency have not been studied.

We propose a computational model of BO-selective cell based on physiology and anatomy of early visual areas, with a focus on the characteristics of latency. In the present study, we examined whether BO-selectivity is evoked from surround modulation that could be formed by feed-forward connections from V1 to V2. Specifically, we carried out Monte Carlo simulations of the model to examine whether the model reproduces the complex behavior of latency, and if so, what characteristics of surround structure are crucial. The results showed that although the behavior of latency is complex, a few, simple anatomical constraints that underlie the surround modulation result in the size and fragment modulation. Specifically, we observed the size and fragment modulations for a small surround at near CRF and a large surround at far. This result supports that feed-forward surround modulation apparent in early visual areas underlies the establishment of BO selectivity.

2 Prediction

Zhang et al. [2] have investigated BO-selective cells in monkey V2 with specific interests on surround structure and latency. They reported two spatial factors in the modulation of latency. First, they investigated how the latency of the BO-selective cell is modulated by the size of square. They observed a short latency (92ms) for a small square (3-4 ^{o} in visual angle), and a long latency (113ms) for a large square (6-8^{o}). Next, they divided a square into eight fragments (four corners and four edges) and presented two fragments, one on the CRF and the other elsewhere outside the CRF, and measured the modulation in latency. They observed a shorter latency for far fragments (91ms) compared to near fragments (107ms). Because the latency should rely critically on the distance of signal transmission, the observation of a shorter latency for a smaller square is natural, but a shorter latency for farther segments appears peculiar. The investigation of these characteristics of latency would lead to understanding the essential mechanisms underlying BO-selectivity.

In the present study, we propose that the latency of BO-selective cells depends on the time necessary for signal transmission through the surround modulation that could transmit signals from far outside the CRF to the cell center presumably on the CRF. First, we consider the physiological phenomenon that a larger square evoked a longer latency (size modulation). The surround region for the BO-selective cell that responded to a stimulus might cover at least a part of the stimulus. As the size of stimulus increased, the distance to the surround region would increase in general, as shown in Figure 1(A). We consider that an origin of the size modulation is the delay of signal transmission from the surround region to the cell center, specifically, the delay depending on the distance between the center of the surround and the CRF of the BO-selective cell. If a surround region is relatively small, the transmission within the surround may be negligible. Size modulation would also be observed in a specific case with a large surround region whose center located near the CRF. In this case, because the surround region covers a whole square, the delay simply depends on the size of the square.

Second, we consider the phenomenon that the location of fragment modulates the latency of BO-selective cells (fragment modulation) when a square is fairly large.

A short latency was observed for a far fragment with respect to the CRF, and a long latency for a near fragment. We consider a large surround region as to include both near and far fragments of a large square. The distance between the *surround center* and a far fragment could be shorter than that between the surround center and a near fragment, as illustrated in Figure 1(B). We propose that the origin of the fragment modulation is the delay of signal transmission from the fragment location to the surround center.

We predict that combinations of these delays between V1 cells, surround center and BO-selective cells underlie the latency characteristics such as the size and fragment modulations observed in the electrophysiology. Size and fragment modulations appear to be puzzled and reciprocal because a shorter latency is observed for a small square but a long latency for near fragment of a large square. A single kind of delay might not sufficient to yield both modulations. It appears the separation of delays (that between the centers of surround and CRF and that within the surround) appears to be a crucial factor. The certain combinations of the location and size of surrounds may also be an important factor. To examine the prediction, we implemented these delays in the model based on physiologically and anatomically realistic data, and carried out the simulations.

3 Model and Methods

We developed a network model of BO-selective cells with a specific focus on the examination of the hypothesis that the delays in the mechanism of surround modulation account for the characteristics of latency observed in BO-selective cells. The model comprised of three layers: retinal layer as an input layer, V1 layer with orientation-selective cells, and V2 layer with BO-selective cells. Figure 2(A) illustrates the outline of the model. To focus on the population behavior of latency, we generated a large number of simplified cells that models the transmission of the spikes originated from V1 cells responding to stimulus contours to BO-selective cells in V2 through the surround modulation.

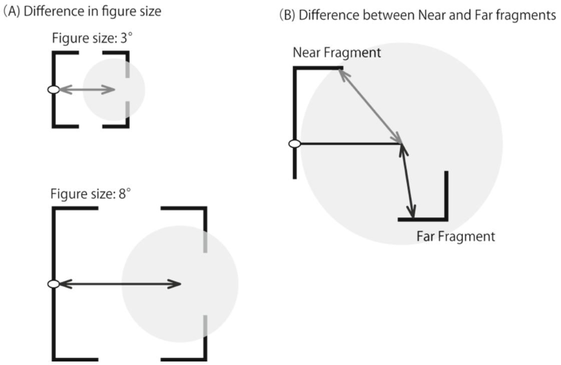

(A) Difference in figure size
Figure size: 3°
Figure size: 8°
(B) Difference between Near and Far fragments
Near Fragment
Far Fragment

Fig. 1. Schematic illustrations to explain the latency characteristics based on signal transmission through surrounding regions. (A) As the size of stimulus increased, the distance to the surrounding region would increase. (B) A large surrounding region should include both near and far fragments. The distance between the surround center and far fragment could be shorter than that between the surround center and a near fragment.

3.1 The Network Model

The model is comprised of three layers. The first layer represents an input image in grey-scale. The second layer consists of orientation-selective cells in V1 arrayed in retinotopy (150x150). The response of a V1 cell is approximated by oriented Gabor filters. The third layer comprised of BO-selective cells with a surround region. BO selectivity is established by surround modulation asymmetric with respect to the CRF [3, 6]. The signals propagated through the surround region modulate linearly the response evoked by a stimulus projected onto the CRF. Because the latency of the signals through the CRF does not vary among the cells, we disregard this pathway in the simulations, and focus on the signal transmission through the surround modulation. The surround regions are circular and defined by its diameter and the distance from the CRF center of the cell. For the sake of simplicity, we focus on BO in the horizontal directions. To test the population behavior, we carried out Monte Carlo simulations with a variety of surround regions.

3.2 The Latency of the Model

We developed a connection model presumably between V1 and V2, as illustrated in Figure 2(A). The delay of signal transmission from V1 cells to BO-selective cells through surround modulation depends on the distance between the location of stimulus fragment and the center of surround region, and the distance between the centers of the surround and the BO-selective cell. The latency of BO signal, , is calculated by:

$$L = L_a + L_b + C \tag{1}$$

, where and are the delays of signal transmission from a stimulus fragment to the surround center, and from the surround center to the BO-selective cell, respectively. We set ms based on [4, 5]. Because we focused on the differences in onset latency between the preferred and non-preferred presentations of stimulus for several stimulus conditions, signal transmissions only through surround modulation were computed given that transmission through the CRF was identical regardless of the presentation side and the stimulus conditions.

3.3 The Simulation Methods

We carried out Monte Carlo simulations to examine whether the model reproduces the physiological observations of the differences in the latency that depends on stimulus conditions. We randomized the size and location of surround regions to examine what characteristics of surround structure are crucial for the behavior. Figure 2(B) shows an overview of simulation settings. The stimulus square was quantized to 2020, with each cell models the response of V1 simple cell. The transmission delay was calculated for the signals originated from each point along the square except those near the CRF. The range of the size and location of surround region were specified in each experiment, with a total of 20,000 regions.

4 Results

We propose that combinations of delays between V1 cells, surround center and BO-selective cells underlie the latency characteristics such as size and fragment modulations observed in the electrophysiology. The crucial question to the proposal is whether this mechanism is valid for numerous combinations of the locations and sizes of surround regions that result in population behavior. To answer this question, we implemented the connections in the model based on physiologically realistic data, and carried out Mote Carlo simulations of the latency with a variety of surround regions based on the signal transmission from every point along the figure contour in V1 to a BO-selective cell in V2 through the surround modulation.

In the first subsection, we show whether in fact the prediction is valid for typical cases in size and fragment modulation, given physiologically realistic constraints. Size and fragment modulations appear to be puzzled and reciprocal because a shorter latency is observed for a small square but a long latency for near fragments of a large square. The surround regions that yield both modulations would hold certain combinations of the location and size of surrounds. In the second subsection, we carry out thorough, exploratory simulations to examine what combinations produce both modulations simultaneously. Finally, we reproduce the population latency, including both size and fragment modulations, with the constraint identified by the second subsection.

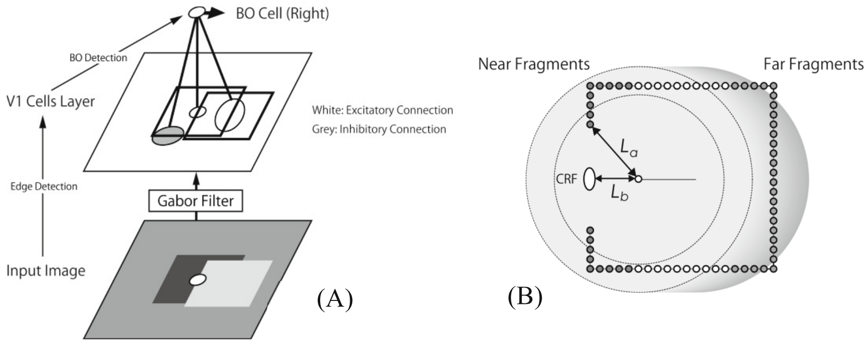


Fig. 2. (A) An illustration of the model architecture consisting of three layers. The layers correspond to (1) stimulus input from the retina, (2) V1 cells, and (3) a BOselective cell in V2. The response of the BO-selective cell is evoked by signals transmitted via connections from V1 cells with the formation of surrounding regions. (B) The spatial arrangement of the model. The latency of a BO-selective cell is calculated based on the distance of signal pathway that originates from the edge of a square through surround center to the center of the BO-selective cell. Small circles indicates the location of the stimulus.

4.1 The Latency Modulation with Specific Surround Regions

We examine whether the prediction is valid for typical cases in size and fragment modulation, given physiologically realistic constraints. There are two mechanisms that would underlie the size modulation. The first is the case with a relatively large surround whose center located near the CRF. With this specific case, the surround region covers a whole square so that the delay depends simply on the size of the square. We carried out Monte Carlo simulations for this case with large surrounds (4^o to 8^o in diameter) centered at near the CRF (0^o to 3^o). A small square and a large square were presented to the model with the size and location of the surround randomized, resulting in a total of 1,520,000 tests (20,000 surrounds x 2 squares x 38 points). Figure 3(A) shows the histograms of the latency for small and large squares. The mean latencies for the small and large squares were 76ms and 97ms, respectively, which reproduces the size modulation. However, it should be noted that this is a specific case valid probably for a small number of cell, because the surround is limited to large and centered near the CRF.

The second mechanism for size modulation that appears to be more general is the delay of signal transmission from the surround region to the cell center, specifically, the delay depending on the distance between the center of the surround and the CRF of the BO-selective cell. If a surround region is relatively small, the transmission within the surround may be negligible. Therefore, size modulation is expected to depend on the location of surround. Here, we test simple models that consist of a single surround region for a single BO-selective cell. When a surround region is fairly small, only a part of the square falls onto the surround. The fragment projected onto the surround is subject to the delay originated from the surround modulation, and other parts of the square are not. If a surround is located far from the cell center (CRF), the delay is evoked only if a large square is presented, and the delay will be fairly long. On the contrary, if a surround is located near the cell center, the delay is evoked if a small square is presented, and it will be rather short. For the sake of simplicity, we compared two cases: a model BO-selective cell with a small surround at near the cell center, and that at far from the center. This corresponds to a

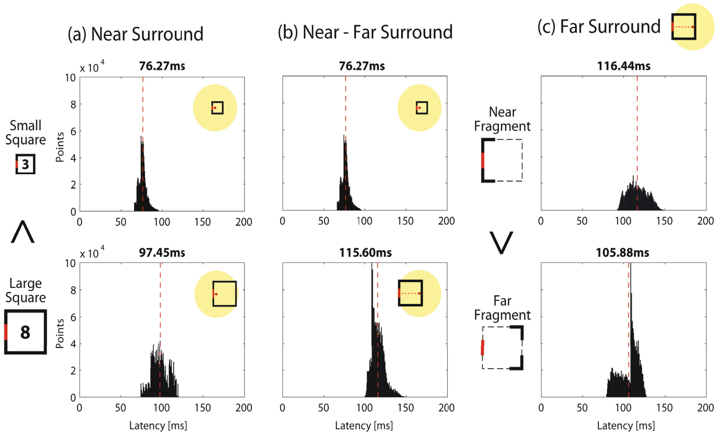


Fig. 3. The histograms of the latency for three typical surround regions. (a&b) The latencies of BO cells with near (a) and near-far (b) surrounds for small (top) and large (bottom) squares. The dotted lines indicate the mean. (c) The latencies of BO cells with far surrounds for near (top) and far (bottom) fragments. The models reproduce the size and fragment modulations.

hypothetical model with two small surround regions, each located near and far from the CRF, respectively. We carried out Monte Carlo simulations of the model with small surrounds (0^o to 4^o in diameter) centered at near (0^o to 3^o) and far (3^o to 6^o) from the CRF, respectively. The histograms of the latencies are shown in Figure 3(B). The mean latencies for the small and large squares were 76ms and 116ms, respectively, reproducing the size modulation. The delay of signal transmission from the surround region to the cell center appears to evoke size modulation if the surround is fairly small.

We propose that fragment modulation is originated from the delay of signal transmission from the fragment location to the surround center. Given a large surround region as to include both near and far fragments of a large square, the distance between the surround center and a far fragment could be shorter than that between the surround center and a near fragment. We carried out Monte Carlo simulations of the model with large surrounds (4^o to 8^o in diameter) centered at far (3^o to 6^o) from the CRF for the fragments of a large square. The histograms of the latencies are shown in Figure 3(C). The mean latencies for the near and far fragments were 106ms and 116ms, respectively, reproducing fragment modulation As we expected, fragment modulation is evoked from the delay of signal transmission from the fragment location to the surround center if a fairly large surround is located far from the CRF.

The simulation results indicate that certain spatial conditions of surround regions yield the size and fragment modulations. In the next subsection, we consider thoroughly what combination of size and location of surround establish both size and fragment modulation.

4.2 The Spatial Characteristics of Surround for Latency Modulation

The previous subsection showed that the latency varies depending on the size and location of surround. In this subsection, we examine what combinations of the size and location of surrounds evoke the size and fragment modulations. We define two indices for the representation of size and fragment modulations:

$$D_{size} = L_{large} - L_{small} \tag{2}$$

$$D_{fragment} = L_{near} - L_{far} \tag{3}$$

, where and are the mean latencies of model cells with specific surround regions in response to *large* and *small* squares, respectively. Similarly, and are the mean latencies of model cells with specific surround regions in response to *near* and *far* fragments, respectively. The surround regions were specified by the size (diameter) and the distance between its center and the CRF. Positive values of the indices indicate the agreement with the size and fragment modulations reported in physiology: a long latency is observed for a large square and a near fragment. Note that the indices were not calculated if either term (e.g., and for) was incomputable (no stimulus fragment falls on to the surround so that no surround modulation was evoked).

Figure 4 shows the computed indices as functions of the size and distance of the surround. Figure 4(A) shows the distribution for the size modulation. Strong positive modulation is observed for large, near surrounds, which agrees with our first hypothesis as described in the previous subsection. We observe negative values for surrounds at very far (7~9^o). As we discussed in the previous subsection, size modulation would not occur when surrounds are located *far* from the CRF because the size modulation depends on the signal transmission between the surround and the cell center. For example, with a large surround at far, signals from contour fragments of a large square (located far from the CRF) travel shorter to reach the surround center (far from the CRF) compared to those from a small square (located near the CRF). This would not be applied for small surrounds on which only either square was projected. Note that small, far surrounds were not computed in this simulation because a small square did not fall onto these surrounds and did not evoked surround modulation.

Figure 4(B) shows the distribution for the fragment modulation. We observe positive modulation for the most of range, but negative modulation for surrounds at very near (0-2^o). As we discussed in the previous subsection, fragment modulation would not occur when surrounds are located near the CRF because the fragment modulation depends on the signal transmission between the location of stimulus fragment and the surround center. Strong positive modulation is observed for large, far surrounds, as expected from the hypothesis. Note that small, far surrounds were not computed in this simulation because near fragments did not fall onto these surrounds and did not evoke surround modulation. As similar to size modulation, negative modulation is expected for these surrounds.

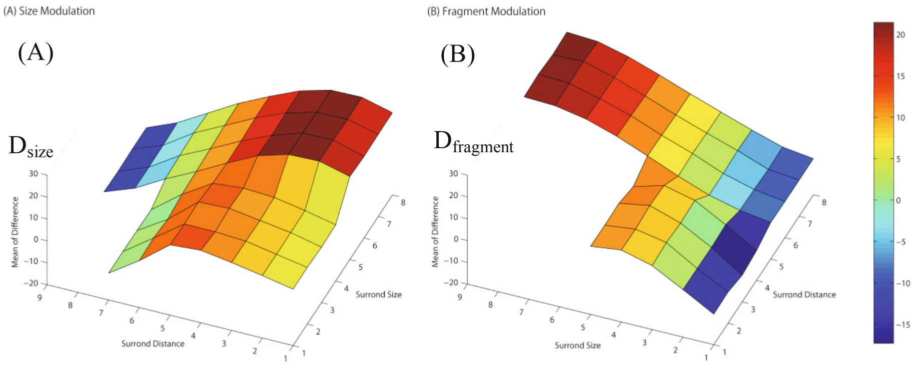


Fig. 4. The computed indices for the size modulation (A) and the fragment modulation (B), as functions of the size (y) and the distance (x) of surround region. The models with near-small and far-large surrounds appear to reproduce the two modulations.

The distributions of the size and fragment modulations showed good agreements with our hypotheses. It is expected that simultaneous establishment of the two modulations requires certain combinations of constraints. It appears that near (1-4^o), small (1-4 ^{o}) regions and far (3-6 ^{o}), large (4-7 ^{o}) regions yield positive values in both size and fragment modulations. In the next subsection, we examine whether such combinations of the size and location of the surround evoke simultaneously the size and fragment modulations.

4.3 Reproduction of the Size and Fragment Modulation

The Monte Carlo simulations have revealed the spatial characteristics of surrounds necessary for the size and fragment modulation of latency, respectively. The results indicate that small, near surrounds and large, far surrounds yield both modulations simultaneously. The combinations of smaller surrounds for nearer locations, and larger surrounds for farer locations appear to be natural from the physiological viewpoint. We carried out the simulations of the model that was comprised of small, near surrounds and large, far surrounds in order to examine whether these surrounds evoke both modulations. Specifically, we set the size and the distance from the CRF of small, near surrounds to 0-4^o and 0-3^o, respectively, and large, far surrounds to 4-8^o and 3-6^o, respectively. The computed latencies of the model were shown in Figure 5. The mean latencies for the small and large squares were 95.2ms and 117.8ms, respectively, reproducing size modulation. The mean latencies for the near and far fragments were 106.1ms and 99.7ms, respectively, reproducing fragment modulation. These latencies show good agreement with those reported in the electrophysiology.

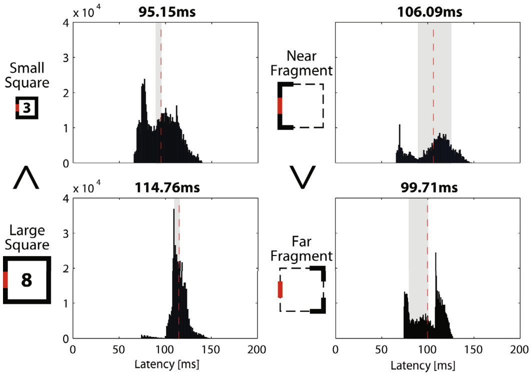


Fig. 5. The latency histograms for BO cells with near-small and far-large surrounds. The conventions the same as Fig. 3. The model reproduces both size and fragment modulations.

5 Discussions

We examined the hypothesis that delays in feed-forward mechanisms of surround modulation account for the characteristics of latency observed in BO-selective cells. Specifically, we proposed that combinations of delays between V1 cells, surround center and BO-selective cells underlie the latency characteristics such as size and fragment modulations observed in the electrophysiology. The crucial question to the proposal is whether this mechanism is valid for numerous combinations of the locations and sizes of surround regions that result in population behavior. To answer this question, we implemented the connections in the model based on physiologically realistic data, and carried out Monte Carlo simulations of the latency with a variety of surround regions based on the signal transmission from every point along the figure contour in V1 to a BO-selective cell in V2 through the surround modulation. The results indicated that BO-selective cells with a smaller surround region at a nearer location from the CRF and those with a larger surround at a farther location reproduce the size and fragment modulations simultaneously. This constrain appears to be natural from the physiological viewpoint. The model explains nicely the puzzled latency of the fragment modulation. These results support the feed-forward mechanism of surround modulation for the neural mechanism underlying the BO selectivity.

Acknowledgement. This work was supported by grant-in aids (KAKENHI) from Japan Society for the Promotion of Science (22300090), and the MEXT, Japan (23135503, 25135704 (Shitsukan)).

References

1. Zhou, H., Freidman, H.S., von der Heydt, R.: Coding of Border Ownership in Monkey Visual Cortex. J. Neurosci. 20(17), 6594–6611 (2000)
2. Zhang, N.R., von den Heydt, R.: Analysis of the Context Integration Mechanisms Underlying Figure–Ground Organization in the Visual Cortex. J. Neurosci. 30(19), 6482–6496 (2010)
3. Sakai, K., Nishimura, H.: Surrounding suppression and facilitation in the determination of border ownership. J. Cognitive Neurosci. 18(4), 562–579 (2006)
4. Alexander, D.M., Wright, J.J.: The maximum range and timing of excitatory contextual modulation in monkey primary visual cortex. Visual Neurosci. 23, 721–728 (2006)
5. Shushruth, S., Ichida, J.M., Levitt, J.B., Angelucci, A.: Comparison of Spatial Summation Properties of Neurons in Macaque 1 and V2. J. Neurophysiol. 102, 2069–2083 (2009)
6. Sakai, K., Nishimura, H., Shimizu, R., Kondo, K.: Consistent and robust determination of border-ownership based on asymmetric surrounding modulation. Neural Networks 33, 257–274 (2012)

Zero-Crossings with the Precedence Effect for Sound Source Localization in Reverberant Conditions

Sung Jun An[1], Rhee Man Kil[2,*], and Byoung-Gi Lee[3]

[1] Prod. Tech. Team, OLED Businesses, Samsung Display Co., Ltd., Asan-City, Korea
[2] College of Information and Communication Engineering, Sungkyunkwan University
2066, Seobu-ro, Jangan-gu, Suwon, Gyeonggi-do, 440-746, Korea
[3] Future IT R&D Laboratory, LG Electronics, Seoul, Korea
sungjun.an@samsung.com, rmkil@skku.edu, bg10.lee@lge.com

Abstract. This paper presents a new method of zero-crossing-based estimation for sound source directions using the spatial cues such as the inter-aural time differences (ITDs) and inter-aural intensity differences (IIDs) in reverberant conditions. The difficulty of estimating sound source directions in reverberant conditions is that the ITDs and IIDs are changed according to reflected sounds. To cope with this problem, the precedence effect in which the preceding auditory signals suppress the lagging signals unless the intensity of signals is not strong enough, is considered. This psychoacoustic phenomenon is believed to play an important role for capturing the less interrupted spatial cues. From this viewpoint, the precedence effect is implemented in zero-crossing-based sound source localization. As a result, the proposed zero-crossing time difference (ZCTD) method with the precedence effect is able to provide the robust estimation of sound source directions even in severely reverberant conditions.

Keywords: sound source localization, zero-crossings, precedence effect, reverberant conditions.

1 Introduction

In the mammalian auditory system, sound source localization relies on the comparison of auditory input obtained from two separate ears. The main cues are inter-aural time differences (ITDs) and inter-aural intensity differences (IIDs). It is widely known that ITDs are the main cues used at low frequencies of less than 1.5 kHz while IIDs are used in the high frequency range [1]. Estimations of sound source directions using ITDs have smaller variations but can be ambiguous at higher frequencies due to phase warping while estimations using IIDs have larger variations but can be used in the high frequency range. However, the ITDs and IIDs are easily affected by noisy and/or reverberant conditions: ITDs can be easily affected by background noise while IIDs can be easily biased

[*] Corresponding author.

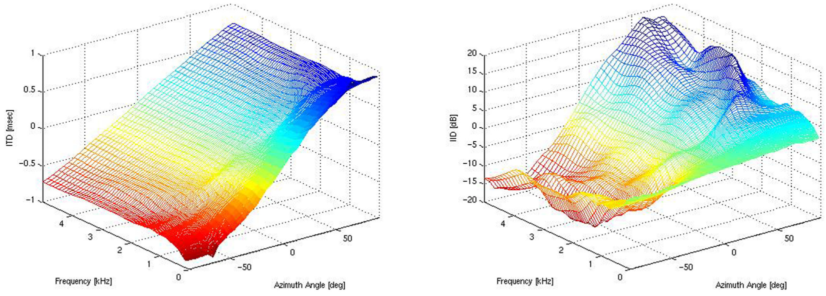
M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 301–308, 2013.

especially in reverberant conditions [2]. In this sense, a reliable and consistent mapping from the spatial cues to the source directions in noisy and/or reverberant conditions should be investigated. For the estimation of ITDs, Jeffress [3] suggested a simple and intuitive hypothesis to measure ITDs in the auditory system. Motivated by Jeffress's model, ITDs are usually estimated using the cross-correlation (CC) of firing rates of auditory signals coming from the channels of left and right ears. However, this approach requires high computational complexity involved in the computation of CC, and they suffer from inaccuracies in estimating the ITDs, especially in noisy multi-source environments. In this context, a method of estimating ITDs using the zero-crossing time differences (ZCTDs) [4] in which the zero-crossings are detected from the filter-bank output of the left and right sensors, was suggested and sucessfully applied to sound source localization in noisy multi-source environments. This method was also successfully applied to the problem of speech source localization, segregation, and recognition for humanoid robots [5]. However, the ZCTD method is not applicable to severely reverberant conditions since the spatial cues such as the ITDs and IIDs are changed due to the reflected sounds under reverberant conditions. In this context, it is important to capture the less interrupted spatial cues. For this purpose, we consider to use the precedence effect [6] in psychoacoustic phenomenon: the preceding auditory signals suppress the lagging signals unless the intensity of signals is not strong enough. It is believed that this effect helps to capture the less interrupted spatial cues. From this viewpoint, the precedence effect is implemented in zero-crossing-based sound source localization. As a result, the proposed ZCTD method with the precedence effect is able to provide the robust estimation of sound source directions even in severely reverberant conditions.

2 Zero-Crossing-Based ITD Estimation

In the estimation of ITDs, it is important to obtain reliable estimates as much as possible. First, as done in the ZCPA coding [7], a series of band-pass filter is applied to each left and right sensor signals. Here, let us denote $x_i(t)$ as the output signal of the ith channel of the filter-bank. The estimation of ITDs is performed separately for each channel. Suppose there are N (upward) zero-crossings, and zero-crossing times are represented by t_n, $n = 1, 2, \cdots, N$ satisfying $x_i(t_n) = 0$. To distinguish the signals generated from the left and right sensors, we use $x_i^L(t)$ and $x_i^R(t)$ as the signal at the ith channel of the left and right sensors, respectively. We now describe the principle of determining the ITD using zero-crossings. Let us define zero-crossing time in the left channel as t_n^L for $n = 1, 2, \cdots, N$ and in the right channel as t_m^R for $m = 1, 2, \cdots, M$ where N and M represent the number of zero-crossings detected from the left and right channel signals, respectively. In the auditory information processing, the ITDs and IIDs convey same information of sound source directions. From this observation, we consider the method of selecting valid ITD-IID sample pairs. First, for t_n^L, we consider the following candidates of ITD estimates:

$$\Delta t_i(n, m) = t_n^L - t_m^R, \tag{1}$$



(a) (b)

Fig. 1. The maps of azimuth angles versus location cues: (a) and (b) represent the maps of azimuth angles versus ITD and IID values, respectively

where t_m^R is selected within a window in which the time interval is given by the range of $t_n^L - T$ and $t_n^L + T$ for the time span T. The time span T is determined as $1ms$. Here, the problem is to determine the proper t_m^R for the given t_n^L. To solve this problem, we consider the following IID estimates:

$$\Delta p_i(n,m) = 10 \log_{10} \frac{p_n^L}{p_m^R}, \tag{2}$$

where p_n^L and p_m^R represent the left power at time n and the right power at time m, respectively. They are defined by

$$p_n^L = \frac{1}{2W} \sum_{t=t_n^L-W}^{t_n^L+W} (x_i^L(t))^2 \quad \text{and} \tag{3}$$

$$p_m^R = \frac{1}{2W} \sum_{t=t_m^R-W}^{t_m^R+W} (x_i^R(t))^2, \tag{4}$$

where W is set to 5/center frequency of the ith channel. Since the ITD and IID estimates represent the same information of sound source direction, we consider to make a map $f_T(\theta)$ of sound source directions represented by the angles measured from the frontal axis (azimuth angles) versus ITD values and also a map $f_I(\theta)$ of azimuth angles versus IID values. These maps can be made by investigating the ITD or IID values for the corresponding sound source angles as illustrated in Fig. 1. Using these maps, we can identify the corresponding angle value θ_{ITD} for the ITD value $\Delta t_i(n,m)$ from the map $f_T(\theta)$; that is, $\theta_{ITD} = f_T^{-1}(\Delta t_i(n,m))$. This inverse mapping is in general possible since in most cases, the map $f_T(\theta)$ is a monotonously increasing function while the map $f_I(\theta)$ is usually not a monotonously increasing functions as illustrated in Fig. 1. Here, we can identify the corresponding IID value for θ_{ITD} from the map $f_I(\theta)$. Then, we can search the best matching ITD-IID sample pair by comparing the

IID value $\Delta p_i(n,m)$ with $f_I(\theta_{ITD})$; that is, we can select the time index k for the proper ITD value as

$$k = \arg\min_m |\Delta p_i(n,m) - f_I(f_T^{-1}(\Delta t_i(n,m)))|. \tag{5}$$

As a result, we get the ITD sample $\Delta t_i(n)$ at the ith channel as

$$\Delta t_i(n) = t_n^L - t_k^R, \tag{6}$$

where k satisfies (5). Then, in the case of no intensity difference between two sensors, the SNR of the filtered signal can be approximated as

$$SNR \approx 10 \log_{10} \frac{1}{w_i^2 Var(\Delta t_i(n))}, \tag{7}$$

where w_i represents the frequency of the filtered signal. Here, the estimated SNR can be effectively used to construct the histogram of ITD samples in noisy environments. However, this method is not applicable to severely reverberant conditions since the spatial cues such as the ITDs and IIDs are changed due to the reflected sounds under reverberant conditions. In this context, the precedence effect is implemented in the previously suggested ZCTD method.

3 ZCTD Algorithm with the Precedence Effect

In this section, we consider the identification of reliable ITD estimates using the precedence effect [6] in which preceding auditory signals suppress the lagging signals unless the intensity of signals is not strong enough. In the human auditory system, the precedence effect can be described as follows: the preceding signal suppresses the lagging signals which 1) arrive within 40 msec after arrival of the leading signal and 2) are not stronger than the leading signal by more than 10-15dB. Let us apply this rule to the zero-crossing-based ITD estimation. As described in the previous section, for each zero-crossing point t_n^L, the ITD estimate $\Delta t_i(n)$ and the signal power p_n^L are obtained. According to the precedence effect, we consider the following rule of selecting unreliable zero-crossing points:

Precedence effect for selecting unreliable ZC points

- The ZC point is determined as an unreliable cue if

$$t_n^L - t_m^L < \delta_t \quad \text{and} \quad 10 \log_{10} \frac{p_n^L}{p_m^L} < \delta_p \quad \text{for } m < n, \tag{8}$$

  where δ_t and δ_p represent the threshold for the precedence time window and the precedence power difference, respectively.
- Otherwise, the ZC point is determined as a reliable cue.

Here, for our method of the ITD estimates, δ_t is set to 40 msec which is similar with the human auditory system and δ_p is set between 0 and 5 dB to obtain the reasonable number of ITD estimates. This rule is applied to the previously mentioned ZCTD method. For the estimation of ITDs using zero-crossings under reverberant conditions, the following algorithm of the ZCTD with the precedence effect is suggested:

ZCTD with the precedence effect

Step 1. (Detection of Zero-Crossings) Zero-Crossings in the upward direction are detected from the filter-bank output of the left and right sensors.

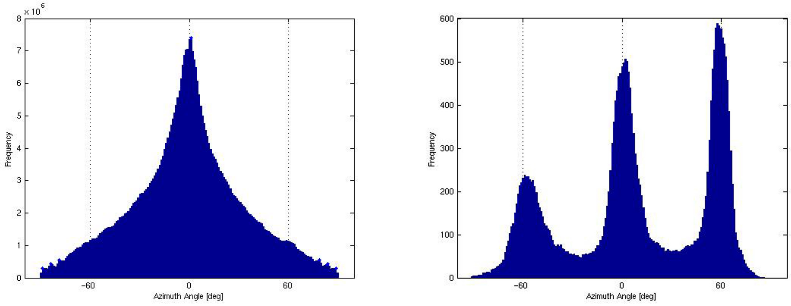
Step 2. (Application of the Precedence Effect) The precedence effect is applied to zero-crossing points of the left and right channels of the filter-bank output. As a result, the reliable and unreliable ZC points are determined from the condition of (8).

Step 3. (Selection of consistent ITD-IID sample pairs) For each ZC point of the left channel, determine valid ITD estimates satisfying the condition of (5).

Step 4. (Collection of reliable ITD estimates) If both zero-crossing points of the ITD estimate; that is, t_n^L and t_k^R for the computation of $\varDelta t_i(n)$ as described in (6), are selected as reliable cues by the precedence effect of Step 2, the ITD estimates are collected; otherwise, the ITD estimates are discarded.

Step 5. (Decision of Sound Source Directions) A Histogram of the collected ITD estimates in Step. 4 is made and the sound source directions are determined by the values of azimuth angles corresponding to the peaks of the histogram. For more detail procedure of selecting the peaks of the histogram, refer to [4].

After the detection of ZCs coming from the filter-bank channels of the left and right sensors, the suggested algorithm selects reliable ITD cues using the precedence effect of (8): that is, by checking whether the ZCs coming from the left and right channels are passed or not by the condition of (8), we can eliminate the unreliable cues of ITD estimates. Then, by accumulating the reliable ITD estimates, the predominant cues which are mainly caused by direct arrivals from the sound sources, are obtained and this provides us to collect reliable ITD estimates which are less interrupted by the reflected sounds. As a result, we can make robust decision of sound source directions even in the reverberant conditions. As an example of the suggested algorithm, the histograms of ITD estimates using the ZCTD method alone and the ZCTD method with the precedence effect in which the threshold for the precedence power difference δ_p was set as 3 dB, are compared as illustrated in Fig. 2. In this example, three sound sources uttered by three male speakers are located at -60, 0, and 60 degrees of azimuth angles under the reverberant condition of a small room with brick walls. In Fig. 2-(a), the histogram of ITD estimates using the ZCTD method showed the peak at the center and decreasing envelope of ITD estimates as the absolute degree of azimuth angle increases. This is a typical histogram of ITD estimates in severely reverberant conditions. This happens when the reflected sounds are coming from every direction of the wall so that there are not much difference between the left and right ITD cues in the majority of ITD estimates. The ITD histogram of



(a) (b)

Fig. 2. The ITD histograms using ZCTD methods under the reverberant condition of a room with brick walls: (a) and (b) represent the ITD histograms using the ZCTD and ZCTD with the precedence effect methods, respectively.

Fig. 2-(b) showed that the suggested ZCTD method with the precedence effect illustrated clear distinction of sound source directions appeared as distinctive local peaks in the ITD histogram. This example demonstrates that the precedence effect is quite effective to filter out these reflected sounds.

4 Simulation

For our simulation of binaural information processing, sound source signals were transformed by the Head-Related-Transfer-Function (HRTF) [8] and decomposed by a gamma-tone filter-bank with 128 channels, in which the center frequencies are linearly spaced in equivalent rectangular bandwidth (ERB) scale from 80 Hz to 5.0 kHz. For the precedence effect, the threshold for the precedence power difference δ_p was set as 3 dB since it provided reasonable selection of reliable ITD estimates in our simulation. Then, for each channel, the direction angles corresponding to the measured ITDs were obtained from the previously made (azimuth angle versus ITD) lookup table. Accumulating these data across channels and for a speech segment, we made a histogram of estimated angles: the direction angles corresponding to the measured ITD samples were collected and the estimated direction angle was determined by the peak value of the histogram. Here, the angle resolution (or bin size) of the histogram is given by 1 degree.

For the simulation of reverberant signals, we use RoomSim impulse response generating tool [9]. By convolving source signals with the impulse responses of RoomSim, we can obtain the binaural sound signals in reverberant conditions. For this simulation, we considered the configuration of our simulation conditions as described in Table 1. For the wall conditions, we considered the anechoic, percent 50, acoustic plaster, plywood, and brick walls. Here, the equivalent reverberation times RT_{60} (according to Sabine's formula [9]) for the percent 50, acoustic plaster, plywood, and brick walls are calculated as 143 ms, 230 ms, 542 ms, and 2,577 ms, respectively. From these results, the percent 50 and acoustic plaster walls can be considered as the mild and medium levels of reverberant walls, respectively while the plywood and brick walls can be considered as the severely reverberant walls.

Table 1. Configuration of simulation conditions for sound source localization

Room Size	Length (x): Width (y): Height (z) = 6m: 4m: 3m
Wall Conditions	Anechoic, Percent 50, Acoustic Plaster, Plywood, and Brick
Ear Position	x = 3.5m, y = 1.5m, z = 1.2m
Head Width	14.5 cm
Distance from a Source	1.5 m

As a benchmark data set, the digit utterances obtained from 10 male and 10 female speakers were selected from the TIDigits in Aurora 2 database. For each speaker, 10 utterances were selected and thus in total, 200 utterances were used for our simulation. Then, the simulation for sound source localization using the ZCTD and ZCTD with the precedence effect (ZCTD_PREC) was performed. Then, the simulation results of root mean square errors (RMSEs) between the true and estimated sound source directions for various azimuth angles of 10, 20, 40, 60 and 80 degrees and also for various wall conditions were obtained as illustrated in Fig. 3. In Fig. 3-(a), the estimation errors of the ZCTD method increased as the degree of azimuth angle increased in severely reverberant conditions. This was due to the fact that the histogram of ITD estimates had the peak at the center of azimuth angle as illustrated in Fig. 2-(a). On the other hand, in Fig. 3-(b), the estimation errors of the ZCTD method with the precedence effect were small and did not have strong tendency of estimation errors according to the azimuth angles. In summary, these simulation results showed that 1) both ZCTD-based methods provided more accurate estimation of sound source directions especially in smaller azimuth angles, 2) the ZCTD method provided robust estimation of sound source directions up to the medium level of reverberant condition such as a room with acoustic plaster walls except high azimuth angles, and 3) the ZCTD_PREC provided quite robust estimation of sound source directions even in the severely reverberant conditions such as a room with plywood walls and a room with brick walls.

5 Conclusion

In this work, the precedence effect in which the preceding auditory signals suppress the lagging signals unless the intensity of signals is not strong enough, is considered. This psychoacoustic phenomenon is believed to play an important role for identifying the directions of sound sources. From this viewpoint, the precedence effect was implemented in the previously suggested method [4] of zero-crossing-based sound source localization. In our approach, the precedence effect is easily implemented in the ZCTD method by adding the condition of

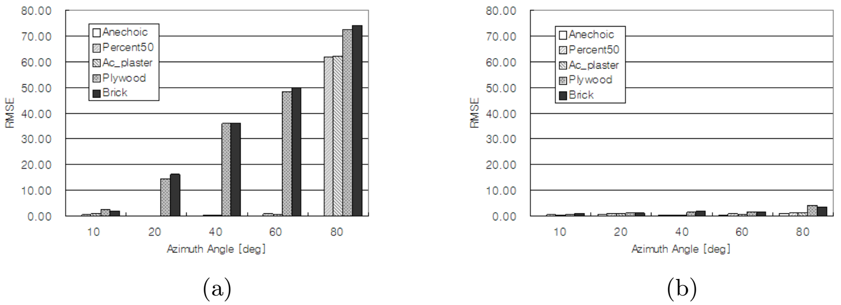


Fig. 3. Sound source localization for a source in various reverberant conditions: (a) and (b) represent the RMSE of the estimated sound source directions in degrees of azimuth angles using the ZCTD and ZCTD with the precedence effect method, respectively.

determining unreliable ZC points. This is possible since the ITD cues are actually measured in our method as does in the human auditory system. As a result, the suggested ZCTD method with the precedence effect is able to provide the robust estimation of sound source directions in reverberant conditions. Through the simulation for sound source localization in various reverberant conditions, we have shown that the suggested ZCTD method with the precedence effect provides the robust estimation of the sound source directions even in severely reverberant conditions.

References

1. Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press, Cambridge (1997)
2. Shinn-Cunningham, B., Kopco, N., Martin, T.: Localizing nearby sound sources in a classroom: Bianural room impulse responses. Journal of Acoustic Society of America 117(5), 3100–3115 (2005)
3. Jeffress, L.: A place theory of sound localization. Journal of Computational Physiology and Psychology 41, 35–39 (1948)
4. Kim, Y., Kil, R.: Estimation of inter-aural time differences based on zero-crossings in noisy multi-source environments. IEEE Transactions on Audio, Speech, and Language Processing 15(2), 734–743 (2007)
5. An, S., Kil, R., Kim, Y.: Zero-crossing-based speech segregation and recognition for humanoid robots. IEEE Transactions on Consumer Electronics 55(4), 2341–2348 (2009)
6. Litovsky, R., Colburn, H., Yost, W., Guzman, S.: The precedence effect. Journal of Acoustic Society of America 106(4), 1633–1654 (1999)
7. Kim, D., Lee, S., Kil, R.: Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on Speech and Audio Processing 7(1), 55–69 (1999)
8. Gardner, W., Martin, K.: HRTF measurement of a kemar. Journal of Acoustic Society of America 97, 3907–3908 (1995)
9. Campbell, D.: ROOMSIM user guide (V3.3) (2004)

Ensemble Joint Approximate Diagonalization by an Information Theoretic Approach

Yoshitatsu Matsuda and Kazunori Yamaguchi

Department of General Systems Studies,
Graduate School of Arts and Sciences, The University of Tokyo,
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan
{matsuda,yamaguch}@graco.c.u-tokyo.ac.jp

Abstract. Joint approximate diagonalization (JAD) is a widely-used method for blind source separation, which can separate non-Gaussian sources without any other prior knowledge. In this paper, a new extension of JAD (named ensemble JAD) is proposed in order to ameliorate the robustness for a small size of samples by an information theoretic approach. In JAD, the cumulant matrices are estimated and represented by the average (namely, the first-order moment) over given samples. On the other hand, ensemble JAD preserves the ensemble of all the cumulant matrices for each sample without averaging them. Then, the second-order moments among the ensemble are utilized for estimating the sources. Numerical experiments verify the validity of this method when the sub-Gaussian (negative kurtosis) sources are included.

Keywords: independent component analysis, joint approximate diagonalization, information theoretic approach, higher-order cumulants.

1 Introduction

Independent component analysis (ICA) is a useful method in signal processing [5,4]. It can solve blind source separation problem under the assumptions that source signals are statistically independent of each other. In the linear model (given as $\boldsymbol{x}^m = \boldsymbol{A}\boldsymbol{s}^m$), it estimates the $N \times N$ mixing matrix $\boldsymbol{A} = (a_{ij})$ and the N-dimensional source signals $\boldsymbol{s}^m = (s_i^m)$ from only the N-dimensional observed signals $\boldsymbol{x}^m = (x_i^m)$. Each m corresponds to a sample ($m = 1, \cdots, M$). So, N and M are the number of signals and the sample size, respectively. Joint approximate diagonalization (denoted by JAD) [3,2] is one of the widely-used methods for estimating $\boldsymbol{A}$. JAD utilizes the following algebraic property of cumulant matrices: $\tilde{\boldsymbol{\Gamma}}_{pq} = \boldsymbol{W}\boldsymbol{\Gamma}_{pq}\boldsymbol{W}'$ is diagonal for any p and q if $\boldsymbol{W} = (w_{ij})$ is equal to the separating matrix $\boldsymbol{A}^{-1}$, where $\boldsymbol{\Gamma}_{pq} = (\kappa_{ijpq})$ and $\tilde{\boldsymbol{\Gamma}}_{pq} = (\tilde{\kappa}_{ijpq})$ are the matrix of the 4-th order cumulants of $\boldsymbol{x}^m$ and the rotated one, respectively. The error function of JAD is defined as the sum of off-diagonal elements $\sum_{p,q>p}\sum_{i,j>i}\left(\tilde{\kappa}_{ijpq}\right)^2$. A significant advantage of JAD is its versatility. It does not depend on the specific statistical properties of sources except for non-Gaussianity [2]. However, because the actual estimation $\boldsymbol{\Gamma}_{pq} = \sum_m \boldsymbol{\Gamma}_{pq}^m / M$ is just an approximation of the true

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 309–316, 2013.

$\boldsymbol{\Gamma}_{pq}$ (where $\boldsymbol{\Gamma}_{pq}^{m}$ is the estimation of cumulants on a single sample m), JAD is not expected to be robust for a small size of samples. In our previous works [6,7], an information theoretic approach to JAD has been proposed in order to improve the robustness. By estimating theoretically the "true" probability distribution of each non-diagonal $\tilde{\kappa}_{ijpq} \in \tilde{\boldsymbol{\Gamma}}_{pq}$ $(i \neq j)$ under the conditions that $\boldsymbol{W}$ is accurately estimated (namely, $\boldsymbol{W} = \boldsymbol{A}^{-1}$), this approach could apply some robust statistical techniques to JAD.

In this paper, the information theoretic approach is developed further and a new robust expansion of JAD is proposed. First, the "true" probability distribution on all the elements of $\tilde{\boldsymbol{\Gamma}}_{pq}^{m} = (\tilde{\kappa}_{ijpq}^{m})$ (including the diagonal ones) for each sample m is estimated theoretically. Second, a new objective function is derived as the log-likelihood of the distribution. Then, a new method called "ensemble JAD" is proposed by optimizing the objective function. This method holds the ensemble of the cumulant matrices $\tilde{\boldsymbol{\Gamma}}_{pq}^{m}$ for all the samples and estimates $\boldsymbol{W}$ by utilizing all the matrices. Though its computational costs increase in proportion to the sample size M, it can utilize the additional information of the limited samples which is lost in the usual JAD through the average operation $\tilde{\kappa}_{ijpq} = \sum_{m} \tilde{\kappa}_{ijpq}^{m} / M$. This paper is organized as follows. In Section 2, the true probability distribution of $\tilde{\kappa}_{ijpq}^{m}$ is estimated theoretically. The new method "ensemble JAD" is described in Section 3. Section 4 shows numerical results on some artificial datasets. Lastly, this paper is concluded in Section 5.

2 Estimation of Distribution

Here, the "true" distribution of $\tilde{\kappa}_{ijpq}^{m}$ is estimated. The word "true" means that the separating matrix is given accurately $\boldsymbol{W} = \boldsymbol{A}^{-1}$. Now, the following four conditions are given in advance:

1. **Linear ICA Model:** The linear ICA model $\boldsymbol{x}^{m} = \boldsymbol{A}\boldsymbol{s}^{m}$ holds. In addition, the mean and the variance of each independent source s_{i}^{m} are 0 and 1, respectively. In other words, $E_{s}(s_{i}^{m}) = 0$ and $E_{s}(s_{i}^{m} s_{j}^{m}) = \delta_{ij}$ where $E_{s}()$ is the expectation operator over $\boldsymbol{s}^{m}$ and δ_{ij} is the Kronecker delta.
2. **Dominance of Kurtosis:** The kurtosis of s_{i}^{m} is dominant over the other higher-order cumulants. In other words, the 3rd-order, 5th-order, and the other higher-order cumulants are negligible.
3. **Whitening:** $\boldsymbol{x}^{m}$ is accurately whitened ($\boldsymbol{A}$ is orthogonal).
4. **Random Mixture:** Each element a_{ij} in $\boldsymbol{A}$ is given randomly and independently, whose mean and variance are 0 and $1/N$, respectively. In other words, $E_{A}(a_{ij}) = 0$ and $E_{A}(a_{ij} a_{kl}) = \delta_{ik}\delta_{jl}/N^{2}$ where $E_{A}()$ is the expectation operator over $\boldsymbol{A}$.

Because Condition 3 constrains $\boldsymbol{A}$, Condition 4 does not hold rigorously under this condition. However, because this constraint is relatively weak, both of Conditions 3 and 4 can be satisfied approximately by generating $\boldsymbol{A}$ randomly and whitening $\boldsymbol{x}^{m}$.

Under the above conditions, the first and second moments of the true distribution of $\tilde{\kappa}_{ijpq}^{m}$ $(i \leq j,\ p < q)$ is estimated. Because $\boldsymbol{x}^m$ is assumed to be pre-whitened (Condition 3), the estimated 4th-order cumulants on $\boldsymbol{x}^m$ are given as

$$\kappa_{ijpq}^{m} = x_i^m x_j^m x_p^m x_q^m - \delta_{ip}\delta_{jq} - \delta_{jp}\delta_{iq}. \tag{1}$$

Under the basic linear model (Condition 1), each element $\tilde{\kappa}_{ijpq}^{m}$ of the diagonalized matrix $\tilde{\boldsymbol{\Gamma}}_{pq}^{m} = \boldsymbol{W}\boldsymbol{\Gamma}_{pq}^{m}\boldsymbol{W}'$ is given as

$$\tilde{\kappa}_{ijpq}^{m} = \sum_{k,l} a_{pk} a_{ql} s_i^m s_j^m s_k^m s_l^m - a_{pi} a_{qj} - a_{pj} a_{qi} \tag{2}$$

where $\boldsymbol{x}^m = \boldsymbol{A}\boldsymbol{s}^m$ and $\boldsymbol{W} = \boldsymbol{A}^{-1} = \boldsymbol{A}'$ are utilized. Then, the expectation of the first moment (the mean) of $\tilde{\kappa}_{ijpq}^{m}$ over $\boldsymbol{s}$ and $\boldsymbol{A}$ is given as

$$E_A\left(E_s\left(\tilde{\kappa}_{ijpq}^{m}\right)\right) = \sum_{k,l} \delta_{pq}\delta_{kl} E_s\left(s_i^m s_j^m s_k^m s_l^m\right) - 2\delta_{pq}\delta_{ij} = 0 \tag{3}$$

where Condition 4 and $p < q$ are used. Next, the second-order moment of $\tilde{\kappa}_{ijpq}^{m}(i \leq j)$ and $\tilde{\kappa}_{klrs}^{m}(k \leq l)$ over $\boldsymbol{A}$ and $\boldsymbol{s}$ is given as

$$\begin{aligned} & E_A\left(E_s\left(\tilde{\kappa}_{ijpq}^{m}\tilde{\kappa}_{klrs}^{m}\right)\right) \\ = & \frac{\delta_{pr}\delta_{qs}}{N^2}\left(\sum_{t,u} E_s\left(s_i^m s_j^m s_k^m s_l^m s_t^m s_t^m s_u^m s_u^m\right) - 4\delta_{ik}\delta_{ij}\delta_{il}\left(E_s\left(\left(s_k^m\right)^4\right) - 1\right)\right) \end{aligned} \tag{4}$$

By a basic relation between cumulants and moments [8], the expectation term over $\boldsymbol{s}^m$ in Eq. (4) is given as a sum of products of cumulants over all the partitions of the set of the subscripts as follows:

$$\sum_{t,u} E_s\left(s_i^m s_j^m s_k^m s_l^m s_t^m s_t^m s_u^m s_u^m\right) = \sum_{t,u} \sum_{\omega \in \Omega_{ijklttuu}} \prod_{B \in \omega} \kappa^s[B] \tag{5}$$

where $\Omega_{ijklttuu}$ is the set of all the partitions of the subscripts $\{i, j, k, l, t, t, u, u\}$, ω is a partition, B is a subset of subscripts in ω, and $\kappa^s[B]$ is the true cumulant of $\boldsymbol{s}^m$ on the subscripts in B. Under Condition 1, the first and second order cumulants κ_i^s and κ_{ij}^s are given as 0 and δ_{ij}, respectively. In addition, the subscripts must be identical in the same subset in a partition because the sources are independent of each other. Therefore, Eq. (5) is rewritten as a combination of the 3rd-8th order cumulants of each s_i^m. Under Condition 2, it is simplified further by focusing on only the 4th-order cumulants (kurtoses, denoted as α_i below). In addition, by the symmetric property, the sets of four given subscripts (i, j, k, l) are classified into the following five types: (i, j, k, l), (i, i, j, k), (i, i, i, j), (i, i, j, j), and (i, i, i, i) where i, j, k, and l are different from each other. Thus, after some lengthy but simple transformations, Eq. (4) is finally rewritten as

$$E_A\left(E_s\left(\tilde{\kappa}_{ijpq}^{m}\tilde{\kappa}_{klrs}^{m}\right)\right) \simeq \begin{cases} \sigma_{ii} & (p = r, q = s, i = j = k = l), \\ \sigma_{ij} & (p = r, q = s, i = k, j = l, i < j), \\ \sigma_{ik} & (p = r, q = s, i = j, k = l, i \neq k), \\ 0 & \text{(otherwise.)} \end{cases} \tag{6}$$

where σ_{vw} is given as

$$\sigma_{vw} = \begin{cases} \alpha_v + 3 + \frac{30\alpha_v+30}{N} + \frac{(\alpha_v+3)\sum_t \alpha_t + 172\alpha_v + 34\alpha_v^2 + 64}{N^2} & (v = w), \\ 1 + \frac{2\alpha_v+2\alpha_w+10}{N} + \frac{\sum_t \alpha_t + 16\alpha_v + 16\alpha_w + 2\alpha_v\alpha_w + 24}{N^2} & (v \neq w). \end{cases} \tag{7}$$

Here, $\boldsymbol{\alpha} = (\alpha_i)$ can be regarded as the unknown parameters. Eqs. (6) and (7) show that the second moments are determined by the unknown parameter $\boldsymbol{\alpha}$ through an $N \times N$ matrix $\Sigma = (\sigma_{ij})$. After all, the first-order moment of each $\tilde{\kappa}^m_{ijpq}$ in Eq. (3) and the second-order moment between any $\tilde{\kappa}^m_{ijpq}(i \leq j)$ and $\tilde{\kappa}^m_{klrs}(k \leq l)$ in Eq. (6) were estimated.

The Gaussian approximation model using the estimated first and second moments is employed in this paper. Regarding $\tilde{\kappa}^m_{jipq}(i < j)$, it is equal to $\tilde{\kappa}^m_{ijpq}$ by algebraic symmetry. Therefore, any fixed prior distribution can be employed without changing the likelihood essentially. An independently and identically uniform distribution $u\,(x) = c$ is used here for simplicity. By noting that the covariance between $\tilde{\kappa}^m_{ijpq}$ and $\tilde{\kappa}^m_{klrs}$ is equal to 0 unless $p = r$ and $q = s$, each set $\tilde{\boldsymbol{\Gamma}}^m_{pq} = \left(\tilde{\kappa}^m_{ijpq}\right)$ (for $i = 1, \ldots, N$ and $j = 1, \ldots, N$) is regarded to be independent of $\tilde{\boldsymbol{\Gamma}}^m_{rs}$. Thus, the "true" distribution of $\tilde{\boldsymbol{\Gamma}}^m_{pq}$ $(p < q)$ is estimated as follows:

$$P^{\tilde{\Gamma}}\left(\tilde{\boldsymbol{\Gamma}}^m_{pq}\right) = c^{N(N-1)/2} \prod_{i,j>i} g^{ij}_{\text{off}}\left(\tilde{\kappa}^m_{ijpq}\right) g_{\text{on}}\left(\text{dg}(\tilde{\boldsymbol{\Gamma}}^m_{pq})\right) \tag{8}$$

where $g^{ij}_{\text{off}}()$ is a Gaussian distribution of each off-diagonal elements $(i < j)$ given by

$$g^{ij}_{\text{off}}(\tilde{\kappa}) = \exp\left(-\tilde{\kappa}^2/2\sigma_{ij}\right)/\sqrt{2\pi\sigma_{ij}} \tag{9}$$

and $g_{\text{on}}()$ is a multidimensional Gaussian distribution of the vector of the on-diagonal elements $\text{dg}(\tilde{\boldsymbol{\Gamma}}^m_{pq})$ by

$$g_{\text{on}}\left(\text{dg}(\tilde{\boldsymbol{\Gamma}}^m_{pq})\right) = \exp\left(\frac{-\text{dg}(\tilde{\boldsymbol{\Gamma}}^m_{pq})'\boldsymbol{\Sigma}^{-1}\text{dg}(\tilde{\boldsymbol{\Gamma}}^m_{pq})}{2}\right)/\sqrt{2\pi|\boldsymbol{\Sigma}|}, \tag{10}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. $c^{N(N-1)/2}$ corresponds to the off-diagonal symmetric elements $(i > j)$. By the transformation $\tilde{\boldsymbol{\Gamma}}^m_{pq} = \boldsymbol{W}\boldsymbol{\Gamma}^m_{pq}\boldsymbol{W}'$, the linear transformation matrix from the vectorized elements of $\boldsymbol{\Gamma}^m_{pq}$ to those of $\tilde{\boldsymbol{\Gamma}}^m_{pq}$ is given as the Kronecker product $\boldsymbol{W} \otimes \boldsymbol{W}$. Therefore, the distribution of $\boldsymbol{\Gamma}^m_{pq}$ is determined by

$$P^{\Gamma}\left(\boldsymbol{\Gamma}^m_{pq}|\boldsymbol{W}, \boldsymbol{\Sigma}(\boldsymbol{\alpha})\right) = |\boldsymbol{W} \otimes \boldsymbol{W}| P^{\tilde{\Gamma}}\left(\tilde{\boldsymbol{\Gamma}}^m_{pq}|\boldsymbol{W}, \boldsymbol{\Sigma}\right) = P^{\tilde{\Gamma}}\left(\tilde{\boldsymbol{\Gamma}}^m_{pq}|\boldsymbol{W}, \boldsymbol{\Sigma}\right) \tag{11}$$

where $|\boldsymbol{W} \otimes \boldsymbol{W}| = |\boldsymbol{W}|^{2N} = 1$ because of the orthogonality of $\boldsymbol{W}$. Thus, the distribution of $\boldsymbol{\Gamma}^m_{pq}$ is estimated, which depends on the parameters $\boldsymbol{W}$ and $\boldsymbol{\Sigma}$. Note that $\boldsymbol{\Sigma}$ is determined by $\boldsymbol{\alpha}$ in Eq. (7).

3 Ensemble Joint Approximate Diagonalization

In order to estimate the optimal parameters of $\boldsymbol{W}$ and $\boldsymbol{\alpha}$, the objective function is defined as the log-likelihood of P^{Γ} in Eq. (11), which is given as

$$\ell\left(\boldsymbol{W}, \boldsymbol{\Sigma}\left(\boldsymbol{\alpha}\right)\right) = \sum_{m,p,q>p} \log P^{\Gamma}\left(\boldsymbol{\Gamma}_{pq}^{m} | \boldsymbol{W}, \boldsymbol{\Sigma}\right)$$

$$= \sum_{m,p,q>p}\left(-\sum_{i,j>i}\left(\frac{\log\sigma_{ij}}{2} + \frac{\left(\tilde{\kappa}_{ijpq}^{m}\right)^{2}}{2\sigma_{ij}}\right) - \frac{\log|\boldsymbol{\Sigma}| + \operatorname{dg}\left(\boldsymbol{\Gamma}_{pq}^{m}\right)' \boldsymbol{\Sigma}^{-1}\operatorname{dg}\left(\boldsymbol{\Gamma}_{pq}^{m}\right)}{2}\right) \tag{12}$$

where some constants are omitted. Because it is difficult to maximize ℓ w.r.t $\boldsymbol{W}$ and $\boldsymbol{\alpha}$ simultaneously, an alternating algorithm is employed where $\boldsymbol{\alpha}$ and $\boldsymbol{W}$ are optimized alternately.

Regarding the optimization of ℓ w.r.t. $\boldsymbol{\alpha}$ for a fixed $\boldsymbol{W}$, a gradient algorithm is employed in this paper. Because every $\tilde{\kappa}_{iipq}^{m}$ is fixed, Eq. (12) is rewritten as

$$\ell\left(\boldsymbol{\Sigma}\left(\boldsymbol{\alpha}\right)\right) = -\frac{\sum_{i,j\neq i}\log\sigma_{ij}}{4} - \frac{\log|\boldsymbol{\Sigma}|}{2} - \sum_{i,j}\frac{v_{ij}}{4\sigma_{ij}} - \frac{\sum_{i,j} z_{ij}\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}}{2} \tag{13}$$

where $\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}$ is the (i,j)-th element of the inverse matrix $\boldsymbol{\Sigma}^{-1}$ and some constants are neglected. $\boldsymbol{V} = (v_{ij})$ and $\boldsymbol{Z} = (z_{ij})$ are given as

$$v_{ij} = \begin{cases} 0 & (i = j), \\ \frac{\sum_{m,p,q>p}\left(\tilde{\kappa}_{ijpq}^{m}\right)^{2}}{MN(N-1)/2} & (i \neq j), \end{cases} \tag{14}$$

and

$$z_{ij} = \frac{\sum_{m,p,q>p}\tilde{\kappa}_{iipq}^{m}\tilde{\kappa}_{jjpq}^{m}}{MN\left(N-1\right)/2}. \tag{15}$$

The gradient $\nabla\ell\left(\boldsymbol{\alpha}\right)$ $(= \frac{\partial\ell}{\partial\alpha_i})$ is given as

$$\frac{\partial\ell}{\partial\alpha_i} = \sum_{p,q}\frac{\partial\sigma_{pq}}{\partial\alpha_i}\frac{\partial\ell}{\partial\sigma_{pq}} \tag{16}$$

where

$$\frac{\partial\sigma_{pq}}{\partial\alpha_i} = \begin{cases} \delta_{pi}\left(1 + \frac{30}{N} + \frac{172+\sum_t\alpha_t+68\alpha_p}{N^2}\right) + \frac{\alpha_p+3}{N^2} & (p = q), \\ \left(\delta_{pi} + \delta_{qi}\right)\left(\frac{2}{N} + \frac{16}{N^2}\right) + \frac{1+2\delta_{pi}\alpha_q+2\delta_{qi}\alpha_p}{N^2} & (p \neq q) \end{cases} \tag{17}$$

and

$$\frac{\partial\ell}{\partial\sigma_{pq}} = -\frac{\left(1-\delta_{pq}\right)}{4\sigma_{pq}} - \frac{\left(\boldsymbol{\Sigma}^{-1}\right)_{pq}}{2} + \frac{v_{pq}}{4\sigma_{pq}^{2}} + \frac{\sum_{k,l} z_{kl}\left(\boldsymbol{\Sigma}^{-1}\right)_{kp}\left(\boldsymbol{\Sigma}^{-1}\right)_{ql}}{2}. \tag{18}$$

Here, the basic property $\mathrm{d}\left(\boldsymbol{C}^{-1}\right)_{ij}/\mathrm{d}c_{kl} = -\left(\boldsymbol{C}^{-1}\right)_{ik}\left(\boldsymbol{C}^{-1}\right)_{lj}$ for any invertible matrix $\boldsymbol{C} = (c_{ij})$ is utilized. Thus, the following update equation is derived by the gradient method:

$$\boldsymbol{\alpha} := \boldsymbol{\alpha} + \tau \nabla \ell \left(\boldsymbol{\alpha}\right). \tag{19}$$

where τ is the stepsize parameter determined by a simple line search.

Regarding the optimization of ℓ w.r.t. $\boldsymbol{W}$ for a fixed $\boldsymbol{\alpha}$ (and $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$), the Jacobi method is employed in the similar way as in the well-known JADE algorithm [3], which repeats a simple rotation of each pair of signals in order to optimize the total objective function. For a pair $(i, j > i)$, ℓ is simplified into the following term ℓ_{ij} depending on i and j:

$$\begin{aligned}\ell_{ij}\left(\theta\right) = &-\sum_{m,p,q>p}\left(\frac{\left(\bar{\kappa}_{ijpq}^{m}\right)^{2}}{2\sigma_{ij}} + \frac{\left(\boldsymbol{\Sigma}^{-1}\right)_{ii}}{2}\left(\bar{\kappa}_{iipq}^{m}\right)^{2} + \frac{\left(\boldsymbol{\Sigma}^{-1}\right)_{jj}}{2}\left(\bar{\kappa}_{jjpq}^{m}\right)^{2}\right)\\ &-\sum_{m,p,q>p}\left(\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}\bar{\kappa}_{iipq}^{m}\bar{\kappa}_{jjpq}^{m}\right)\end{aligned} \tag{20}$$

where $\bar{\kappa}_{ijpq}^{m}$ is the rotated cumulant by θ on the pair (i, j). They are given by

$$\begin{pmatrix}\bar{\kappa}_{iipq}^{m}\left(\theta\right) & \bar{\kappa}_{ijpq}^{m}\left(\theta\right)\\ \bar{\kappa}_{ijpq}^{m}\left(\theta\right) & \bar{\kappa}_{jjpq}^{m}\left(\theta\right)\end{pmatrix} = \begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\tilde{\kappa}_{iipq}^{m} & \tilde{\kappa}_{ijpq}^{m}\\ \tilde{\kappa}_{ijpq}^{m} & \tilde{\kappa}_{jjpq}^{m}\end{pmatrix}\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}. \tag{21}$$

Therefore, $\ell_{ij}\left(\theta\right)$ is rewritten as

$$\ell_{ij}\left(\theta\right) = \beta_{1}\sin 4\theta + \beta_{2}\cos 4\theta + \beta_{3}\sin 2\theta + \beta_{4}\cos 2\theta + \beta_{5} \tag{22}$$

where

$$\beta_{1} = \frac{\gamma_{1}}{4}\sum_{m,p,q>p}\left(\tilde{\kappa}_{iipq}^{m}\tilde{\kappa}_{ijpq}^{m} - \tilde{\kappa}_{ijpq}^{m}\tilde{\kappa}_{jjpq}^{m}\right), \tag{23}$$

$$\beta_{2} = \frac{\gamma_{1}}{16}\sum_{m,p,q>p}\left(\left(\tilde{\kappa}_{iipq}^{m}\right)^{2} + \left(\tilde{\kappa}_{jjpq}^{m}\right)^{2} - 2\tilde{\kappa}_{iipq}^{m}\tilde{\kappa}_{jjpq}^{m} - 4\left(\tilde{\kappa}_{ijpq}^{m}\right)^{2}\right), \tag{24}$$

$$\beta_{3} = \frac{\gamma_{2}}{2}\sum_{m,p,q>p}\left(\tilde{\kappa}_{iipq}^{m}\tilde{\kappa}_{ijpq}^{m} + \tilde{\kappa}_{jjpq}^{m}\tilde{\kappa}_{ijpq}^{m}\right), \tag{25}$$

$$\beta_{4} = \frac{\gamma_{2}}{4}\sum_{m,p,q>p}\left(\left(\tilde{\kappa}_{iipq}^{m}\right)^{2} - \left(\tilde{\kappa}_{jjpq}^{m}\right)^{2}\right), \tag{26}$$

$$\gamma_{1} = -\frac{1}{\boldsymbol{\Sigma}_{ij}} + \left(\boldsymbol{\Sigma}^{-1}\right)_{ii} + \left(\boldsymbol{\Sigma}^{-1}\right)_{jj} - 2\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}, \tag{27}$$

$$\gamma_{2} = \left(\boldsymbol{\Sigma}^{-1}\right)_{ii} - \left(\boldsymbol{\Sigma}^{-1}\right)_{jj}, \tag{28}$$

and β_{5} is a constant. The optimal $\hat{\theta}$ is given as

$$\hat{\theta} = \mathrm{argmax}_{\theta}\ell_{ij}\left(\theta\right). \tag{29}$$

Though the optimum of $\ell_{ij}\left(\theta\right)$ is not given analytically, it is not difficult to maximize numerically such a periodic function with a single parameter. A simple

MATLAB function "fminbnd" is employed in this paper. Regarding the termination condition for each pair optimization, a model selection approach using AIC is employed in the same way as in [6]. Consequently, the condition is given as the following inequality (the details are described in [6]):

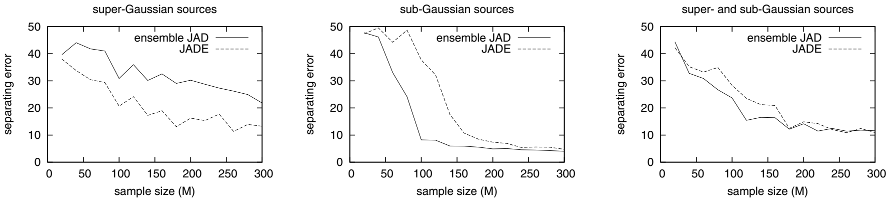
$$\ell_{ij}\left(\hat{\theta}\right) - \ell_{ij}\left(0\right) > 1. \tag{30}$$

The complete algorithm of ensemble JAD is given as follows:

1. *Initialization.* Whiten given observed signals $\boldsymbol{x}^m$ and multiply it by a randomly generated orthogonal matrix. Then, calculate every $\boldsymbol{\Gamma}^m_{pq}$.
2. *Optimization (A).* Calculate $\boldsymbol{V}$ and $\boldsymbol{Z}$, and estimate $\hat{\boldsymbol{\alpha}}$ by repeating Eq. (19) until convergence.
3. *Optimization (B).* For every pair (i, j),
   (a) Calculate $\hat{\theta}$ by Eq. (29).
   (b) Only if Eq. (30) is satisfied, rotate $\boldsymbol{W}$ and every $\tilde{\boldsymbol{\Gamma}}^m_{pq}$ by $\hat{\theta}$.
4. *Convergence decision.* If no pair has been rotated in Optimization (B), end. Otherwise, go back to Optimization (A).

4 Results

Ensemble JAD was compared with the well-known JADE [3] in blind source separation of some artificial sources. The number of sources N was set to 8. The following three types of sources were used: (a) all the sources were generated by the Laplace distribution (super-Gaussian); (b) all the sources by the uniform distribution (sub-Gaussian); (c) the half of them (4 sources) were generated by the Laplace one and the other half by the uniform one. The square mixing matrix $\boldsymbol{A}$ was randomly generated. The size of samples M was set from 20 to 300. The final errors at the convergence were measured by the medians of Amari's separating error [1] over 10 runs. Fig. 1 shows the curves of the final separating errors of ensemble JAD and JADE along the size of samples. Ensemble JAD was superior to JADE when the sources include sub-Gaussian ones. Ensemble JAD could estimate a more accurate separating matrix for a smaller size of samples than JADE. Especially when all the sources were sub-Gaussian, ensemble JADE could achieve the accurate estimation by about half size of samples in comparison with JADE. On the other hand, ensemble JAD was inferior to JADE when all the sources are super-Gaussian. Through close inspection of this inferiority, it was found that the kurtoses in $\boldsymbol{\alpha}$ were not estimated accurately in this case. The estimation may be improved in the future by assuming more general conditions in the theoretical estimation of the true distribution and employing more sophisticated optimization methods for $\boldsymbol{\alpha}$. Nevertheless, the numerical results verified that the current ensemble JAD is effective for a limited size of samples at least when the sources include sub-Gaussian ones.



(a) super-Gaussian. (b) sub-Gaussian. (c) super- and sub-Gaussian.

Fig. 1. Comparison of the final separating errors along the size of samples: The solid and dotted curves correspond to ensemble JAD and JADE, respectively

5 Conclusion

In this paper, we propose a new ICA method named ensemble JAD, which utilizes all the cumulant matrices on the ensemble of samples instead of the averaged estimation. Numerical results verified that the method was superior to the usual JADE for a limited size of samples including sub-Gaussian sources. We are planning to elaborate the method further and apply it to other datasets. We are also planning to investigate the effects of other higher-order cumulants in addition to kurtoses.

References

1. Amari, S., Cichocki, A.: A new learning algorithm for blind signal separation. In: Touretzky, D., Mozer, M., Hasselmo, M. (eds.) Advances in Neural Information Processing Systems 8, pp. 757–763. MIT Press, Cambridge (1996)
2. Cardoso, J.F.: High-order contrasts for independent component analysis. Neural Computation 11(1), 157–192 (1999)
3. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non Gaussian signals. IEE Proceedings-F 140(6), 362–370 (1993)
4. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. Wiley (2002)
5. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)
6. Matsuda, Y., Yamaguchi, K.: An adaptive threshold in joint approximate diagonalization by assuming exponentially distributed errors. Neurocomputing 74, 1994–2001 (2011)
7. Matsuda, Y., Yamaguchi, K.: A robust objective function of joint approximate diagonalization. In: Villa, A.E.P., Duch, W., Érdi, P., Masulli, F., Palm, G. (eds.) ICANN 2012, Part II. LNCS, vol. 7553, pp. 205–212. Springer, Heidelberg (2012)
8. Mendel, J.: Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. Proceedings of the IEEE 79(3), 278–305 (1991)

EM Training of Hidden Markov Models for Shape Recognition Using Cyclic Strings

Vicente Palazón-González, Andrés Marzal, and Juan M. Vilar[*]

Dept. Llenguatges i Sistemes Informàtics and Institute of New Imaging Technologies
Universitat Jaume I de Castelló. Spain
{palazon,amarzal,jvilar}@lsi.uji.es

Abstract. Shape descriptions and the corresponding matching techniques must be robust to noise and invariant to transformations for their use in recognition tasks. Most transformations are relatively easy to handle when contours are represented by strings. However, starting point invariance is difficult to achieve. One interesting possibility is the use of cyclic strings, which are strings with no starting and final points. Here we present the use of Hidden Markov Models for modelling cyclic strings and their training using Expectation Maximization. Experimental results show that our proposal outperforms other methods in the literature.

Keywords: hidden markov models, cyclic strings, shape recognition.

1 Introduction

In a shape classifier, shapes can be represented by their contours or by their regions. Contour based descriptors are widely used as they preserve local information, which is important in the classification of complex shapes.

Dynamic Time Warping (DTW) is being increasingly applied for shape matching [1]. A DTW-based dissimilarity measure is a natural option for optimally aligning contours, since it is able to align parts as well as points and it is robust to deformations. Hidden Markov Models (HMMs) [2] are also used for shape modelling and classification [3–7]. HMMs have some of the properties of DTW matching and they also provide a probabilistic framework for training and classification.

Shape descriptors, combined with shape matching techniques, must be invariant to many distortions, including scale, rotation, noise, etc. Most of these distortions are relatively easy to deal with. However, invariance to the starting point is difficult to achieve no matter the representation. In the case of HMMs there are several approaches for dealing with this invariance [3, 5–7, 4], but all of them have drawbacks. The best solution to this problem is to consider every possible starting symbol of the string that represents the contour, that is, to use cyclic strings. HMMs can only generate ordinary strings and not cyclic strings. To overcome this problem, in this paper, we present the use of HMMs for modelling cyclic strings and their training using Expectation Maximization. Preliminary work on this problem appears in [8].

[*] Work partially supported by the Spanish Government (TIN2010-18958) and the Generalitat Valenciana (*Prometeo*/2010/028).

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 317–324, 2013.

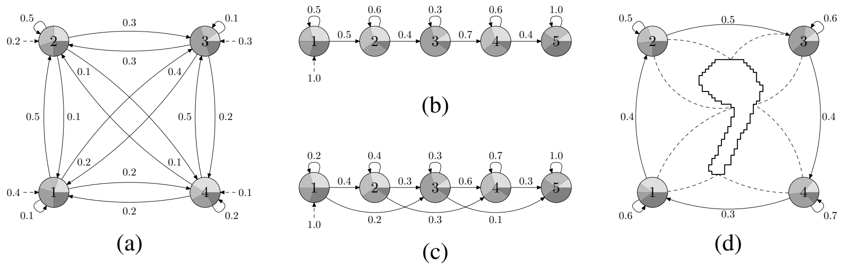


Fig. 1. Examples of HMMs. An HMM state can emit any of four symbols according to the probability distribution represented by a pie chart. Probability distribution for initial states is represented by dotted arrows. (a) Ergodic topology with four states. (b) Linear left-to-right topology. (c) Bakis left-to-right topology. (d) Circular topology as proposed in [4]. Contours are segmented by associating a state to each segment. Ideally, each state is responsible for a single segment.

2 Defining the Problem of Cyclic Strings with HMMs

2.1 Common Approaches

One crucial aspect when applying HMMs is the definition of an adequate topology. Many works use ergodic topologies [5–7], which have some problems. The main is that it is possible to visit a state more than once without using self transitions (Fig. 1a). Ergodic models do not impose restrictions in the order of the strings of observations. When the string of observations is temporal or an order exists (as in shape contours), these topologies do not fully exploit the sequential or temporal information of the data and many states are used to explain multiple observations from different parts along the contour. This makes training and recognition a complex problem.

From the previous observations, left-to-right topologies seem more suitable. These topologies do not allow to visit states that are to the left of the current one (Figs. 1b and 1c). In left-to-right models there is an initial state and a final state. This way, the sequence of states is forced to begin in the initial state and it never revisits a state once it lefts it. When a string of symbols is segmented, all the symbols of a segment are emitted by the same state, and consecutive segments are associated to consecutive states. Although these topologies usually have more states, their number of transitions is low, and consequently the overall complexity of the algorithms is reduced.

In [4], a circular topology is proposed to model contours (Fig. 1d), which can be seen as a modification of the left-to-right topology, where the last emitting state is connected to the first. This topology eliminates the need for a starting point: the contour can be segmented by associating consecutive states to consecutive segments in the cyclic strings, but there is no assumption about which is the first or last segment (Fig. 1d); therefore, there is an analogy with left-to-right topologies. However, there is a problem that breaks this analogy: like in the case of ergodic models all the states can be reached from any state and we can finish in any of them. Therefore, the optimal path can contain non-consecutive repeated states and one state can be responsible of the emission of several non-consecutive segments of the contour. Besides, it is possible to have an optimal path that does not visit all the states at least once.

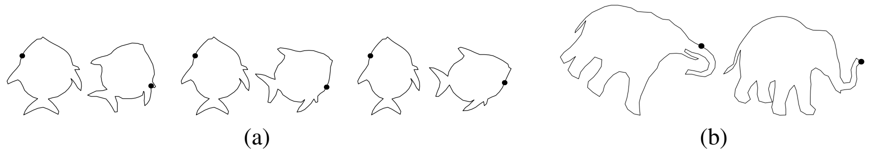

(a) (b)

Fig. 2. (a) Original shapes and their canonical versions using Fourier descriptors [9]. The same shapes compressed in the horizontal axis have different rotation and starting point. (b) Canonical version of an elephant with its trunk down and with its trunk raised using the method of least second moment of inertia [3].

Another approach is the election of a reference rotation and from it a starting point [3, 9]. The basic idea is that, after normalization, all shapes have a canonical version with a "standard" rotation and starting point, and thus, they can be compared as if their representations were linear. But invariance is only achieved for different rotations and starting points of the same shape. Different shapes (even similar ones) may differ substantially in their canonical orientation and starting point. Fig. 2a shows three perceptually similar figures whose canonical versions are significantly different in terms of orientation and starting point. This problem frequently appears in shapes whose basic ellipse is almost a circle. Besides, shapes of the same class with little differences can substantially alter the selection of the starting point. Fig. 2b shows two elephants, one with its trunk down and the other with its trunk raised, this fact and other little differences modify the canonical rotation of the method of least second moment of inertia, and with it, the selection of the starting point.

2.2 Cyclic Strings

The most suitable solution for obtaining the invariance to the starting point is to use every possible starting point of the strings, i.e., using cyclic strings. A cyclic string can be seen as the set of strings obtained by cyclically shifting a conventional sequence. Let $x = x_1 \ldots x_m$ be a string from an alphabet Σ. The cyclic shift $\rho(x)$ of a string x is defined as $\rho(x_1 \ldots x_m) = x_2 \ldots x_m x_1$. Let ρ^k denote the composition of k cyclic shifts and let ρ^0 denote the identity. Two strings x and x' are cyclically equivalent if $x = \rho^k(x')$, for some k. The equivalence class of x is $[x] = \{\rho^k(x) : 0 \leq k < m\}$ and it is called a *cyclic string*.

To achieve starting point invariance using cyclic strings we model the generation process as follows. An HMM generated a string that later suffered an unknown cyclic shift. That is, a model, λ, has generated a string, $x = x_1 x_2 \ldots x_m$, that has suffered the transformation, $\rho^{k'}(x)$, for an unknown k'. We treat x as a cyclic string, $[x]$, and we assume that all the cyclic shifts are equiprobable. Thus, the probability of $[x]$ given a model λ is

$$P([x]|\lambda) = \sum_{k=0}^{m-1} P(x|\lambda, k)P(k|\lambda) = \frac{1}{m} \sum_{k=0}^{m-1} P(\rho^k(x)|\lambda), \tag{1}$$

that is, we must compute the probability for every possible cyclic shift and add them.

Similarly, finding the best cyclic shift and sequence of states amounts to compute

$$\hat{P}([x]|\lambda) = \frac{1}{m} \max_{0\leq k\leq m-1} \hat{P}(\rho^k(x)|\lambda) \propto \max_{0\leq k\leq m-1} \hat{P}(\rho^k(x)|\lambda). \qquad (2)$$

where $\hat{P}$ is the Viterbi score for a string.

Initially, we adopt $\hat{P}$ as an estimation of the real probability because it is a very good approximation[1].

3 Cyclic Training

To train HMMs cyclic strings we have to estimate the Markov model parameters that maximize the probability of the observed cyclic strings. That is, our objective is to maximize:

$$P(X|\lambda) = \prod_{l=1}^{L} P([x]^{(l)}|\lambda) = \prod_{l=1}^{L} \frac{1}{m^{(l)}} \sum_{k=0}^{m^{(l)}-1} P(\rho^k(x^{(l)})|\lambda), \qquad (3)$$

where X is a set of cyclic strings, $X = \{[x]^{(1)}, [x]^{(2)}, \ldots, [x]^{(L)}\}$.

We use an iterative procedure. First, we set some initial values for λ. Then, we obtain new values of these parameters in each iteration, using increasing transformations, applying the Baum-Eagon inequality [10, 11]. It is guaranteed that the new estimated values increase the value of the objective function and, therefore, its convergence.

Let $\Sigma = \{v_1, v_2, \ldots, v_w\}$, be an alphabet (the set of observable events is discrete and finite). Let $A = \{a_{ij}\}$ be the matrix of transition probabilities between states ($1 \leq i, j \leq n$, where n is the number of states) and let $B = \{b_{ij}\}$ be the matrix of observation probabilities ($1 \leq i \leq n$ and $1 \leq j \leq w$) [2]. As we know that $\sum_{j=0}^{n} a_{ij} = 1$ for $0 \leq i \leq n$ and that (3) is a polynomial with respect to A, the new estimation, $\bar{a}_{ij}$, can be obtained with the Baum-Eagon inequality [10, 11]. Applying logarithms and [10] to (3), we conclude that:

$$\bar{a}_{ij} = \frac{\sum_{l=1}^{L}\sum_{k=0}^{m^{(l)}-1} \frac{1}{\sum_{k=0}^{m^{(l)}-1} P(\rho^k(x^{(l)})|\lambda)} \sum_{t=0}^{m^{(l)}-1} \alpha_t^{l_k}(i) a_{ij} b_j(\rho^k(x_{t+1}^{(l)}))\beta_{t+1}^{l_k}(j)}{\sum_{l=1}^{L}\sum_{k=0}^{m^{(l)}-1} \frac{1}{\sum_{k=0}^{m^{(l)}-1} P(\rho^k(x^{(l)})|\lambda)} \sum_{t=0}^{m^{(l)}-1} \alpha_t^{l_k}(i)\beta_{t+1}^{l_k}(j)}, \qquad (4)$$

where $\alpha_t^{l_k}(i)$ and $\beta_t^{l_k}(j)$ are the forward and backward probabilities for the shifted string, $\rho^k(x^{(l)})$, and are defined similarly to those in [2].

Analogously, with $b_i(v_j)$ (the probability of observing the symbol v_j being in state i) and knowing that $\sum_{j=0}^{w} b_i(v_j) = 1$ for $1 \leq i \leq n$ and that (3) is a polynomial with respect to B (the matrix of observation probabilities [2]), we arrive at

$$\bar{b}_i(v_j) = \frac{\sum_{l=1}^{L}\sum_{k=0}^{m^{(l)}-1} \frac{1}{\sum_{k=0}^{m^{(l)}-1} P(\rho^k(x^{(l)})|\lambda)} \sum_{\substack{t=1 \\ \forall \rho^k(x_t^{(l)})=v_j}}^{m^{(l)}-1} \alpha_t^{l_k}(i)\beta_t^{l_k}(i)}{\sum_{l=1}^{L}\sum_{k=0}^{m^{(l)}-1} \frac{1}{\sum_{k=0}^{m^{(l)}-1} P(\rho^k(x^{(l)})|\lambda)} \sum_{t=1}^{m^{(l)}-1} \alpha_t^{l_k}(i)\beta_t^{l_k}(i)}. \qquad (5)$$

[1] And it offers certain implementation advantages and reduction of computational time.

A similar reasoning can be used for continuous models.

This cyclic Baum-Welch (CBW) algorithm is an application of Expectation Maximization, and then, it needs a good initialization for λ. In the next section we propose a heuristic to solve this.

4 A Heuristic for Selecting the Starting Point

We will see in the experiments that the labelling of the training samples gives us an heuristic for obtaining a starting point that improves those in the literature (Section 2).

We need to perform a preprocessing. For it we use cyclic DTW (CDTW) [1] that, apart from returning the cost (distance) of the cyclic alignment, can also return the corresponding cyclic shift of one of the strings for the alignment with the other string. Starting from a set of training samples, our aim is to choose an appropriate starting point for them. We select a representative (the centroid of the class using CDTW) and an arbitrary starting point for it. With the representative of each class and its starting point, we compute the CDTW for each one of the other members of the class and the representative, obtaining the cyclic shift of the alignment that defines a good starting point for each of them. Once we have an appropriate starting point for the training samples, we can train the model of each class as if the cyclic strings were ordinary strings.

In a similar way, to classify a new sample, we begin by finding adequate starting points for it (one for each class). These starting points are computed by CDTW with the representative of each class. Thus, with this starting point for each class we can compute probabilities (or Viterbi scores) in a conventional way.

Although, as we will see in the next section, this solution has worse results than the CBW algorithm, both training and recognition are much faster. Moreover this training can be used as a good initialization for the CBW algorithm.

5 Experiments

In order to assess the behaviour of the presented methods, we performed comparative experiments on a shape recognition task on publicly available databases: MPEG7 CE-Shape-1 corpus part B (MPEG7B) (1440 samples in 70 classes) [12], Silhouette corpus (1070 samples in 41 classes) [13], He-Kundu corpus (8 samples) [3] and Subset 1 corpus (7 classes of the MPEG7B corpus, 20 samples per class) [7].

The outer contours of the images were extracted as sequences of points. A random starting point in each sequence was also selected and 128 landmark points were sampled uniformly. As it is customary in the literature of HMMs and shape recognition we used the curvature shape descriptor. The evaluation was done with classification rates for different number of states (we train an HMM for each class): 10 to 120 in steps of 10. We only use a gaussian per state. The experiments of Section 5.1 were performed using cross validation. The ones of Section 5.2 were performed with a leaving-one-out approach for comparing with other results in the bibliography. In all of them, for classification, we use the Viterbi scores.

5.1 Invariance to the Starting Point, Left-to-right Topologies and Cyclic Approach

In Section 2 several solutions to the starting point invariance problem are commented. We compare our heuristic (Section 4) with the circular topology [4], the election of the starting point using Fourier descriptors [3], and the ergodic topology [5–7]. In Fig. 3a the results of the comparison are shown, for MPEG7B and Silhouette corpora. The election of the starting point and the circular topology (especially the latter) happen to be the most competitive with respect to our heuristic while the ergodic topology obtains the worst results[2].

In Section 2 we mentioned that left-to-right topologies are the most suitable for modelling strings. However, within these topologies, the linear topology seems to be the best for this purpose, because having more transitions increases the complexity of the model. Here we empirically prove this affirmation with a comparison between three left-to-right topologies: linear, Bakis and the one with four transitions per state. The last one is similar to Bakis but with another transition to the state three places from the current. The method used for training and classifying is our heuristic. The results are shown in Fig. 3b. As we can see the linear topology outperforms the others.

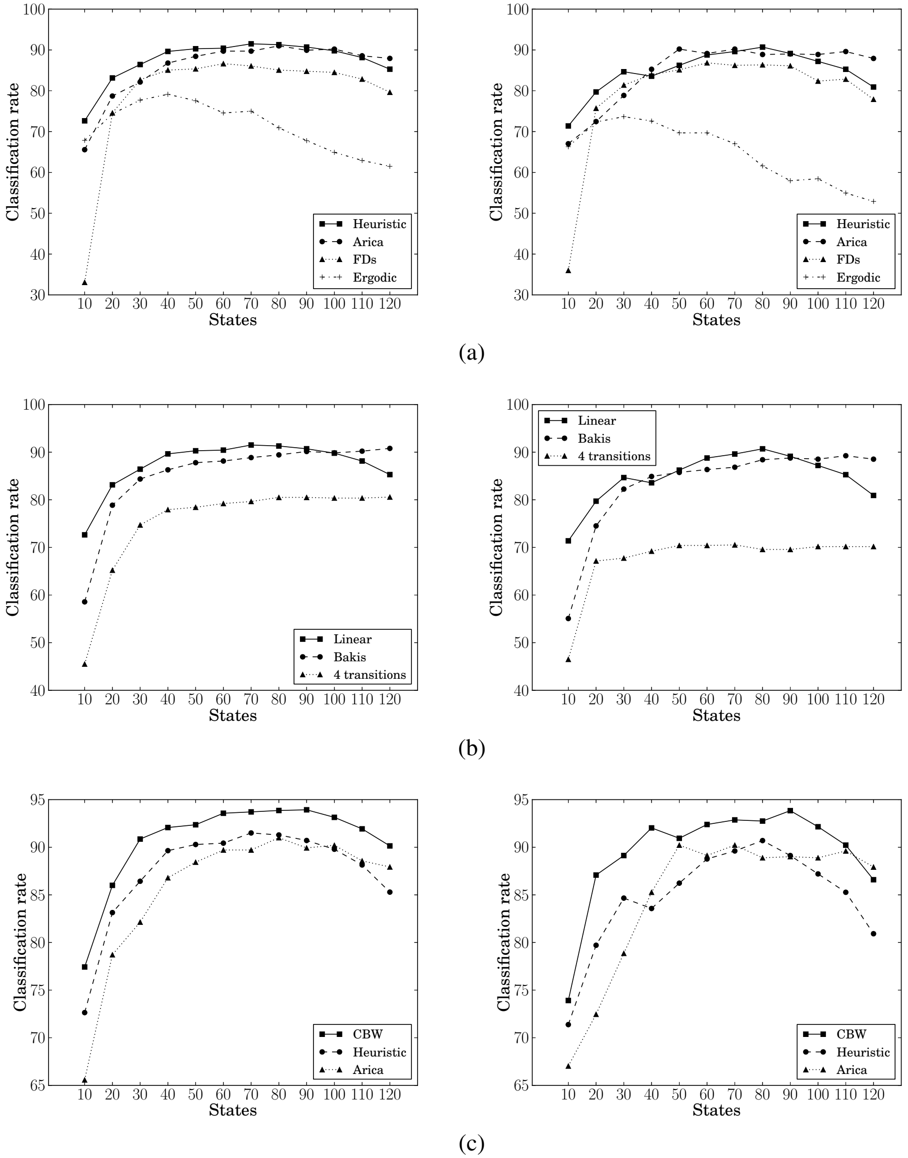
We compare our cyclic approach, the CBW algorithm (Section 3) with our heuristic and the circular topology [4]. Cyclic training is initialized using the heuristic in a linear left-to-right topology. Comparative results are in Fig. 3c. We can observe that cases where CBW algorithm wins predominate. The best results that we obtain using CBW are 93.93% and 93.84% for the MPEG7B and Silhouette corpora.

5.2 More about the Ergodic Topology

In Section 5.1 we experimentally saw that the ergodic topology does not offer good results. However, in the literature there are works [3, 5–7] where this topology is used.

More specifically, in [5] experiments are performed with this topology. For training, the authors choose a number of states with BIC (*Bayesian Inference Criterion*) over a clustering of curvatures. The obtained results are good enough but their corpora have few samples and classes. They use a subset of the MPEG7B corpus of 6 classes with 10 samples per class (a subset of subset 1). They also use He-Kundu corpus for performing an experiment of invariance to the starting point achieving a classification rate of 100%. This way, they conclude that HMMs with an ergodic topology are enough for obtaining this invariance. In our opinion, this experiment is not enough for claiming that affirmation. For this corpus we also achieve a 100% with the CBW algorithm. In [6], a work of the same authors, another subset of MPEG7B is used (a subset of subset 1, with 12 samples per class). We call this corpus subset 2, as it is done in [7]. In this case, they use a canonical method for the election of the starting point. Instead of using BIC for obtaining the number of states, they use a fixed number of states. In [7], the authors, parting from the work of [5, 6], try to improve their results with a training based on GPD (*Generalized probabilistic descent method*). They also use the subset 2 and create a new one, subset 1. With subset 2 they obtain a classification rate of 97.63% ([6] obtains a

[2] We will talk more about this topology in Section 5.2.



(a)

(b)

(c)

Fig. 3. Classification rates (with corpora MPEG7B (left) and Silhouette (right)) for the comparison between: (a) The circular topology (Arica), the election of the starting point with Fourier descriptors (FDs), the ergodic topology (Ergodic) and our heuristic (Heuristic); (b) Different left-to-right topologies. Linear topology, Bakis topology and topology of four transitions. Our heuristic is used for training and classifying; and (c) cyclic Baum-Welch (CBW), our heuristic (Heuristic) and circular topology (Arica).

98.8%). With subset 1 they obtain a 96.43%. With subset 1 and the CBW algorithm, we achieve a 99.28%, that even outperforms the classification rate of [6] with subset 2. None of the previous works show results with the entire MPEG7B corpus.

6 Discussion

In this work, we have argued and empirically proved that other proposals in the literature for obtaining the invariance to the starting point do not offer a suitable solution. We have formalized training and recognition for cyclic strings with Hidden Markov Models, formulating the corresponding Baum-Welch or Expectation Maximization algorithm. We have shown that this cyclic treatment is the current best solution for obtaining the starting point invariance.

References

1. Palazón-Gonzláez, V., Marzal, A.: On the dynamic time warping of cyclic sequences for shape retrieval. Image and Vision Computing 30(12), 978–990 (2012)
2. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77(2) (1989)
3. He, Y., Kundu, A.: 2-D shape classification using hidden Markov model. IEEE Trans. Pattern Anal. Mach. Intell. 13(11), 1172–1184 (1991)
4. Arica, N., Yarman-Vural, F.: A shape descriptor based on circular hidden Markov model. In: ICPR, vol. I, pp. 924–927 (2000)
5. Bicego, M., Murino, V.: Investigating hidden Markov models' capabilities in 2D shape classification. IEEE Trans. Pattern Anal. Mach. Intell. 26(2), 281–286 (2004)
6. Bicego, Murino, Figueiredo: Similarity-based classification of sequences using hidden Markov models. Pattern Recognition 37, 2281–2291 (2004)
7. Thakoor, N., Gao, J., Jung, S.: Hidden Markov model-based weighted likelihood discriminant for 2-D shape classification. IEEE Trans. Image Processing 16(11), 2707–2719 (2007)
8. Palazón, V., Marzal, A., Vilar, J.M.: Cyclic linear hidden Markov models for shape classification. In: Mery, D., Rueda, L. (eds.) PSIVT 2007. LNCS, vol. 4872, pp. 152–165. Springer, Heidelberg (2007)
9. Bartolini, I., Ciaccia, P., Patella, M.: WARP: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance. IEEE Trans. Pattern Anal. Mach. Intell. 27(1), 142–147 (2005)
10. Baum, L.E., Eagon, J.A.: An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model of ecology. Bull. Amer. Math. Soc. 73, 360–363 (1967)
11. Levinson, S.E., Rabiner, L.R., Sondhi, M.M.: An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. The Bell System Technical Journal 62(4), 1035–1074 (1983)
12. Latecki, L., Lakämper, R., Eckhardt, U.: Shape descriptors for non-rigid shapes with a single closed contour. In: CVPR, June 13-15, pp. 424–429. IEEE, Los Alamitos (2000)
13. Sharvit, D., Chan, J., Tek, H., Kimia, B.B.: Symmetry-based indexing of image databases. In: Workshop on Content-Based Access of Image and Video Libraries, pp. 56–62 (1998)

Top-Down Biasing and Modulation for Object-Based Visual Attention

Alcides Xavier Benicasa[1], Marcos G. Quiles[2],
Liang Zhao[3], and Roseli A.F. Romero[3]

[1] Federal University of Sergipe, Itabaiana, Brazil
alcides@{ufs,icmc.usp}.br
[2] Federal University of São Paulo, São Paulo, Brazil
quiles@unifesp.br
[3] University of São Paulo, São Carlos, Brazil
{zhao,rafrance}@icmc.usp.br

Abstract. This work presents a new object-based visual attention model with bottom-up and top-down features. Bottom-up attention is related to the contrast of primitive visual features, such as color, orientation, and intensity. On the other hand, top-down attention is related to the intentions of the viewer and can be seen as a modulation process through the selection system. Thus, if the viewer is searching for an specific shape or color, the top-down modulation can bias the searching process in relation to those features. Our model is composed of five main modules which are responsible for the extraction of the visual features, image segmentation, object recognition, object-saliency map, and object selection. Results on natural images are compared with state-of-the-art approaches and an ground truth fixation maps for a variety of images revealing the efficacy of the proposed approach for visual attention.

Keywords: top-down biasing, bottom-up and top-down visual attention, object-based attention, recognition of objects.

1 Introduction

Perceiving a complex scene is a quite demanding task for a computer albeit our brain does it efficiently. Evolution has developed ways to optimize our visual system in such a manner that only important parts of the scene undergo scrutiny at a given time. This selection mechanism is named visual attention [9].

Visual attention operates in two modes: bottom-up and top-down. Bottom-up attention is driven by scene-based conspicuities, such as the contrast of colors, orientation, etc. [12]. On the other hand, top-down attention is controlled by task, memory, expectations, etc. The knowledge of the task is crucial in this selection mechanism. Top-down attention can even modulate the bottom-up mechanism biasing features according to the task [12].

In additional to modulation mechanism taken into account, what is selected from the scene also represents an important part of the selection process.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 325–332, 2013.

In this scenario, several theories have been proposed and can be gathered in three main lines: location-based attention, feature-based attention, and object-based attention [20].

Object-based models, instead of only delivering the attention to locations or specific features of the scene, claim that the selection if performed on object level, it means that the objects are the basic unit of perception. In this case, since the attention is directed to any part of an object, other parts also benefit from this attentional process [20].

In order to develop models following object-based theories, one needs to consider the integration of a perceptual organization module. This module might segment the objects from the background of the scene based on Gestalt grouping principles, such as similarity connectedness, etc. Those objects, or percepts, will compete for attention [9]. It is worth noting that although these processes can be characterized as bottom-up, they can also be influenced by top-down modulations according to task, such as searching for a specific shape [20].

Several neurocomputational object-based models of visual attention have been proposed in recent years [15,2,3]. Research in qualitative models of visual attention has mainly focused on the bottom-up guidance of early visual features, disregarding any information about objects [12,17]. Achanta et al. [1] proposed a frequency-tuned method that directly defines pixel saliency using a pixel´s color difference from the average image color. Cheng et al. [8] proposed a regional contrast based saliency extraction algorithm, which simultaneously evaluates global contrast differences and spatial coherence. On the other hand, recently works have been conducted regarding the use of the knowledge of the target to influence the computation of the most salient region [14,10,5]. This knowledge is usually learned in a preceding training phase but might in simpler approaches also be provided manually by the user [11]. According to Borji et al. [6], the research in this area is rather new and the few existing models are in their early phases.

Here, we propose a new visual selection model with both bottom-up and top-down modulations. The proposed model integrates a new network-based high level classification [16] and the top-down modulations come in the form of a polarization process in which scene features are biased according to the task. We provide a qualitative comparison of the proposed model against an ground truth fixation maps [13] and two state-of-the-art proposed methods [1,8] for object-based salient region detection.

2 Proposed Model Description

The proposed approach to select salient objects is composed of the following modules: a visual feature extraction module, a top-down biasing feature-based, a LEGION network for image segmentation, a network-based high level data classifier for object recognition, a network of integrate and fire neurons, which creates the object-saliency map and, finally, an object selection module, which highlights the most salient objects in the scene.

The first stage in our visual attention model is responsible for extracting the early visual features in parallel across the scene. The results from this stage are the following conspicuity maps: colors, intensity and orientation. For a complete description of how the conspicuity maps are computed, see [12]. The next stage of the model is the combination of the results from the conspicuity maps with specific weights, for the top-down biasing of the LEGION segmentation network. The implementation of the LEGION followed the algorithm proposed in [18]. The output from those modules feed the following modules: the network for object recognition and the network of integrate and fire neurons, which creates the object-saliency map.

The top-down biasing proposed in our model is defined by the association of weights to output from the conspicuity maps (C_k). The saliency values for all conspicuity maps are weighted and combined into a saliency map S_m defined as:

$$S_m = \frac{1}{n^k} \sum_k W_k C_k, \quad k \in \{color, intensity, orientation\} \tag{1}$$

where n^k denotes the number of conspicuity maps that have been chosen for the biasing and W_k determines weight of the conspicuity map C_k.

According to [18], the segmentation process in the LEGION is based on the idea that a segment must contain at least one oscillator, denoted as a *leader*, which lies in the center of a large homogeneous region. Leaders are all oscillators i in which the lateral potential $p_i \geq \theta$ where θ is a threshold [18]. In order to generate the top-down biasing of the proposed model, an oscillator i defined as leader only will pulse if its saliency value $S_{m_i} \geq \theta_{bias}$.

As described in previous sections, the proposed model takes both bottom-up and top-down modulations into account. Early visual features, i.e. color contrast, intensity contrast and orientation contrast, define the bottom-up signal. On the other hand, information about previously memorized objects and their features (top-down modulation) is responsible for guiding the selection process. Thus, in order to apply the proposed model to select the salient objects of a given scene, the Network-Base High Level Classification (HLC) [16], must be trained with a set of objects representing the desired targets of the scene.

After the training process, HLC network is able to recognize a set of segments (objects). Thus, the overall dynamics of the system can be understood as follows. Each time a segment is highlighted (pulsing) into LEGION network it is directly presented to HLC network. The output of the HLC indicates whether or not the object is among those memorized by the recognition system. If the object is recognized, the output value of the network is used for setting the attribute recognition parameter $R_{i,j}$, where i and j represent the spatial position of pixels inside each segment. Initially, $R_{i,j} = 0$ for all neurons. At the end of this process, all the neurons related to objects that should receive attention (*top-down* modulation) will be assigned to a recognition value ($R_{i,j} = [0, 1]$) that will modulate the attentional process. Segments representing unknown objects can also present nonzero recognition values. In order to avoid those objects receiving top-down modulation, a threshold for the recognition value (θ_r) is adopted. Thus, segments below this threshold are not considered. Hence, the value of recognition $R_{i,j}$ is defined by:

$$R_{i,j} = \begin{cases} 1, \text{ if } R_{i,j} \geq \theta_r \\ 0, \text{ otherwise} \end{cases}, \tag{2}$$

About the prior selection of salients objects, for each pixel of the input scene, the following descriptors were extracted: intensity contrast I, spatial difference in colors RG and RY, orientations O_θ with $\theta \in \{0°, 45°, 90°, 135°\}$, pixel location $[i, j]$, and the recognition value $R_{i,j}$ set by the HLC recognition module.

Let l be a neuron belonging to an active segment into LEGION model, and k its respective index of the features, denoted here as $l_k = [C, I, O, L, R]$ (*Color, Intensity, Orientation, Location, Recognition*). Once the segmentation process is completed and the value of the salience of all pixels that belong to the input image are properly calculated, the average of saliency of each feature k of the segment j can be defined as:

$$S_k^j = \frac{1}{n^j} \sum_{i \in n^j} l_{k_i}^j, \tag{3}$$

where n^j represents the number of neurons in segment j and $l_{k_i}^j$ is the value of the saliency map at neuron i belonging to the feature k at segment j. It is worth to note that each object is still represented by k features, preserving the information of the segment.

According to [19], another important feature that might guide the deployment of attention is the size of the object, which, in this work, is represented by n^j, i.e. the object size is incorporated into the saliency value S_k^j according to the number of neurons representing the segment j. Therefore, the feature vector is redefined and normalized as $l_k = [C, I, O, L, S, R] \in [0, 1]^6$.

It is noted that the segments may not have significant values of saliency. With that in mind, the use of a threshold value of saliency (θ_s) is indicated. Hence, segments with value of saliency below the threshold value won't participate in the competition for attention. The prior selection of salients objects S^j is given by:

$$P_{priorselection}(S^j) = \begin{cases} 1, \text{ if } \left(\frac{1}{n^k} \sum_{i \in n^k} S_i^j\right) \geq \theta_s \\ 0, \qquad \text{otherwise} \end{cases}, \tag{4}$$

where n^k denotes the number of features responsible for guiding the attention considered in this work.

The Object-Salience Map (OSM) is usually defined as a network composed of neurons [15,2,3]. In this work, we do not consider each neuron individually, but groups of neurons that represent objects (Equation (3)). Therefore, the object-salience map is defined as a network composed of objects, with two types of connections: excitatory and inhibitory. Excitatory connections represent a cooperative mechanism responsible for synchronizing groups of objects that closely represent patterns of similarities (objects with similar features). On the other hand, the inhibitory connections are designed to inhibit objects related to background objects of the scene, allowing the object with the greatest saliency within the scene to be selected. When an object pulses in the LEGION, its saliency signal is compared to all the other selected objects, whose states are updated by:

$$\dot{v}_i = -v_i + E_i - W_Y Y_i + \sum_{k=1}^{6} W_k S_k^i, \qquad i = 1, \ldots, n^s \tag{5}$$

where n^s is the number of pulsating objects. The variable v_i represents some voltage-like state of segment i, E_i is the excitatory coupling term and W_Y is the weight of inhibition from the coupling inhibitory Y_i. Equation (5) represents an integrate and fire neuron. Let S_k^j be an object belonging to an active segment into LEGION model, and k its respective index of the features. The excitatory coupling term E_i and the inhibitory coupling term Y_i are defined as it follows:

$$E_i = Y_i = d(S_k^i, S_k^j), \tag{6}$$

in which E_i will be updated if and only if the value of E_i contains the maximum value of activation of the object i, S_k^i represents each pulsating object and k is the feature index. The similarity between the features representing the object S_k^j to the other objects is defined by:

$$d(m, l) = \exp\left(-\sqrt{\sum_{k=1}^{6} W_k (m_k - l_k)^2}\right), \tag{7}$$

where W_k defines the weight associated with each feature k. By adjusting the weights W_k it is possible to direct the attention to desired features. Thus, if $W_k = 0$ for all the primitive information of the input image, the proposed model becomes a strictly top-down model, and if $W_k = 0$ for information related to object recognition, it becomes a bottom-up model.

The inhibitory connections are determined based on the contrast among features. Thus, if two objects share similar features, that is, the contrast between them is small or zero, the term Y_i, in Equation (6), approaches 1.0 (see Equation (7)) and, therefore, the inhibitory coupling term Y_i in Equation (5) assumes a high value. On the other hand, when the signal of such objects is defined by different features, the inhibitory coupling among them is small or even zero. Consequently, objects with similar characteristics are mutually inhibited because of the competition generated by inhibitory connections. An object that has a high contrast in relation to the others is not inhibited and remains oscillating. Thus, it represents the object under focus of attention of the system.

3 Experimental Results

In this section we show how our visual selection model can be used for both feature-based top-down biasing and object-based visual attention. In order to obtain a qualitative comparison of saliency results, we use ground truth fixation maps (FM) for a variety of images publicly available from a database provided by Judd et al. [13], Toronto data set [7] (Fig. 1), and some images proposed by us (Fig. 2). Also depicted are the outputs of both Achanta et al.[1] (AC) and Cheng et al. [8] (CH) algorithms for comparison.

The first set of simulations applies the knowledge of the target to modulate the segmentation. In Fig. 1 three scenes are shown with high variation in color, intensity and orientation. It is worth noting that modifying the parameter values W_{col}, W_{int}, W_{ori}, and θ_{Bias}, can significantly change the number of segments that will participate in the competition for attention. Although it is possible to set the saliency threshold value θ_s, the threshold value of biasing θ_{bias} is important to reduce the computational costs, to maximize detection rate and object recognition. One can say that this behavior is biologically plausible, because in some situations, low-contrast objects are automatically excluded from the scene.

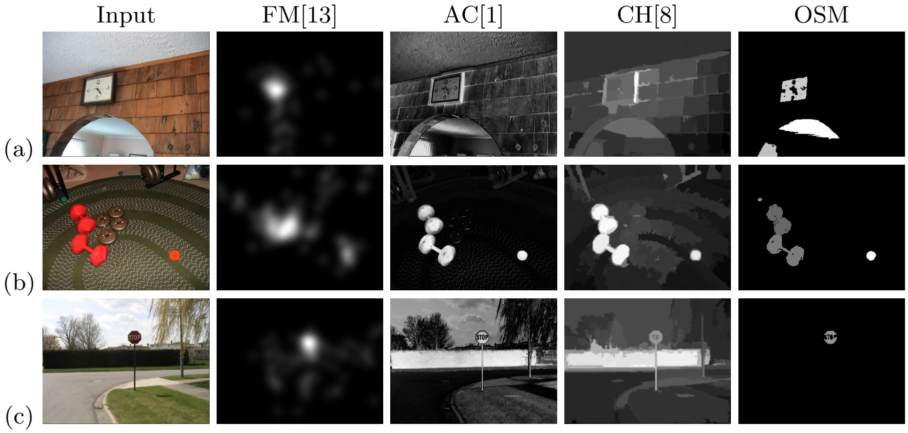


Fig. 1. Results 01 for qualitative comparison. Top-down biasing feature-based is used with the following parameter values: **(a)** $W_{col} = 1.0$, $W_{int} = 1.0$, $W_{ori} = 1.0$, $\theta_{bias} = 0.2$, $W_1 = 1.0$, $W_2 = 0.1$, $W_3 = 0.6$, $W_4 = 0.1$, $W_5 = 0.1$ and $W_6 = 0.0$, **(b)** $W_{col} = 1.0$, $W_{int} = 0$, $W_{ori} = 0$, $\theta_{bias} = 0.1$, $W_1 = 0.3$, $W_2 = 0$, $W_3 = 0$, $W_4 = 0.3$, $W_5 = 0$ and $W_6 = 0$, and **(c)** $W_{col} = 1.0$, $W_{int} = 0$, $W_{ori} = 0$, $\theta_{bias} = 0.5$, $W_1 = 1.0$, $W_2 = 0$, $W_3 = 0$, $W_4 = 0$, $W_5 = 0$ and $W_6 = 0$. The parameter values $\theta_s = 0$ and $W_Y = 1.0$ are used for all the experiments.

In the following simulations, we considered a driver on a road. Here, it is intended to show that the proposed model can indeed help to solve real-world problems, specifically those pertaining to both visual attention and pattern recognition. In this context, initially the HLC network is trained to recognize the road signs shown in Fig. 2 (see [16] for details about this classifier). Then when a segment (neuron) is pulsing in the LEGION network, it is presented to HLC network and its output indicates the classification membership value of each segment. In this simulation, values have been assigned to the parameter in order to direct the attention to recognized objects that are the road sign, since the others were previously excluded by the top-down bias. It is worth noting that in this simulation, due to the similarity among images and goals, it was possible to determine the most appropriate choices of the parameter values. Thus, according to Fig. 2 (OSM), the objects with the greatest saliency value are the road signs.

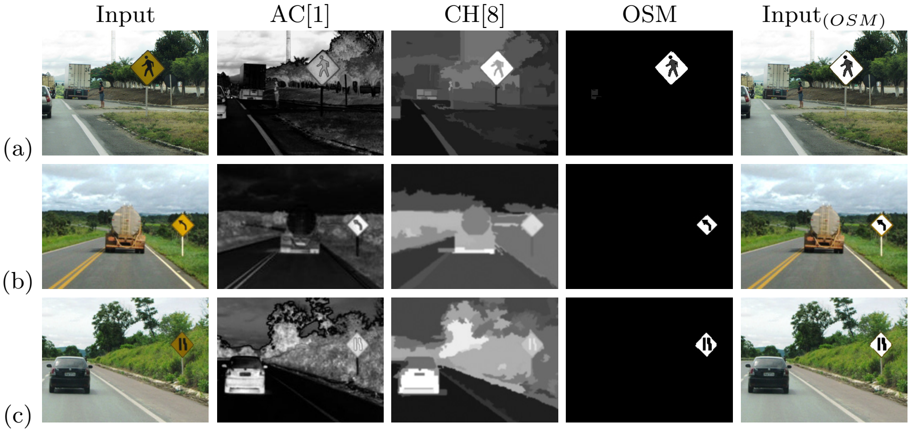


Fig. 2. Results 02 for qualitative comparison. Top-down biasing for road sign detection. The parameter values $W_{col} = 1.0$, $W_{int} = 0.0$, $W_{ori} = 0.0$, $\theta_{bias} = 0.2$, $\theta_s = 0.0$, $W_Y = 1.0$, $W_1 = 1.0$, $W_{[2..5]} = 0$ and $W_6 = 1.0$ are used for all the experiments.

The object-salience map appearing in Figs. 1 and 2 give some sense of the efficacy of the proposed model in predicting which objects of a scene human observers tend to fixate. As may be observed, the predictive capacity of the model is on par with the methods for salient region detection proposed by Achanta et al.[1] and Cheng et al. [8].

4 Conclusion and Future Works

In this work we have presented a new neurocomputational visual attention model with bottom-up selection and top-down modulation. In contrast to former models [4,2,3], the proposed model is able to attend to objects related to the viewers intention. More specifically, by adjusting the weights related to any of the primitive visual features or to the recognition value (higher level intention), it is possible to bias the selection process through the scene. Its dynamics is closed related to biological vision, where top-down modulation can influence the selection of objects related to the task. For example, if we are searching for a red object, the color feature can be biased against other features. Thus, the irrelevant information of the scene according to the task can be ignored. Simulations and qualitative comparisons showed that our model was able to deal with real scenes providing good results in all scenarios taken into account. As future work, we intend to apply the proposed model on video, for this purpose we will consider the top-down bias, able to direct visual attention previously to the segmentation process.

References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned Salient Region Detection. In: IEEE CVPR, pp. 1597–1604 (2009)
2. Benicasa, A.X., Zhao, L., Romero, R.A.F.: Model of top-down / bottom-up visual attention for location of salient objects in specific domains. In: IEEE IJCNN (2012)
3. Benicasa, A.X., Quiles, M.G., Zhao, L., Romero, R.A.F.: An object-based visual selection model with bottom-up and top-down modulations. In: SBRN (2012)
4. Benicasa, A.X., Romero, R.A.F.: Localization of salient objects in scenes through visual attention. In: SBRN (2010)
5. Borji, A., Ahmadabadi, N.M., Araabi, B.N.: Cost-sensitive learning of top-down modulation for attentional control. MVA 22(1), 61–76 (2011)
6. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: A benchmark. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 414–429. Springer, Heidelberg (2012)
7. Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision 9(3), 1–24 (2009)
8. Cheng, M., Zhang, G., Mitra, N.J., Huang, X., Hu, S.: Global contrast based salient region detection. In: IEEE CVPR, pp. 409–416 (2011)
9. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222 (1995)
10. Elazary, L., Itti, L.: A bayesian model for efficient visual search and recognition. Vision Research 50(14), 1338–1352 (2010)
11. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM TAP 7(1), 1–6 (2010)
12. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience 2, 194–203 (2001)
13. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. MIT Computer Science and AI (2012)
14. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimal object detection. In: IEEE CVPR (2006)
15. Quiles, M.G., Wang, D., Zhao, L., Romero, R.A.F., Huang, D.-S.: Selecting salient objects in real scenes: An oscillatory correlation model. Neural Networks 24(1), 54–64 (2011)
16. Silva, T.C., Zhao, L.: Network-based high level data classification. IEEE Transactions on Neural Networks 23, 954–970 (2012)
17. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks 19(9), 1395–1407 (2006)
18. Wang, D., Terman, D.: Image segmentation based on oscillatory correlation. Neural Computation 9, 805–836 (1997)
19. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nature R. Neuroscience 5, 495–501 (2004)
20. Yantis, S.: Goal-directed and stimulus-driven determinants of attentional control, vol. 18, pp. 73–103. MIT Press, Cambridge (2000)

Hidden Markov Model for Action Recognition Using Joint Angle Acceleration

Sha Huang and Liqing Zhang

MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems,
Department of Computer Science and Engineering, Shanghai Jiao Tong University,
Shanghai, 200240, China
shasha8874@sina.com, zhang-lq@cs.sjtu.edu.cn

Abstract. This paper proposes a recognition method of human actions in video by adding new features, the joint angle acceleration to the feature space. In this method, human body is described as three-dimensional skeletons. The features consist of vectors of several important joint angles on the human body and the joint angle accelerations are also considered as a part of features. Hidden Markov Model (HMM) is used as classification scheme. The HMM models are trained by sequences extracted from the CMU graphics lab motion capture database[1]. This method is invariant to scale, coordinate system and transition. A system is implemented to recognize 4 different types of actions (walk, run, jump and jumping jack) both on the dataset of CMU and Weizmann[7]. Each video clip contains a single action type. The experimental results show excellent performance of the proposed approach. A maximum 10.3% accuracy gain can be achieved by our method compared with the method without considering acceleration.

Keywords: Human action recognition, skeleton, joint angle, HMM, angle acceleration.

1 Introduction

Recognition of human actions in video is an important topic in computer vision. It has the wide range of applications, such as security surveillance, video analysis, robot simulation. Even though there are already a number of models and algorithms that can achieve good results under well constraints, the performance of the existing models is not satisfactory for open testing data. For this reason, robust action recognition methods, against the variations of illumination, viewpoint, scale, etc., are demanding for practical applications. Most of the previous methods focus on non-human representations for recognition. That means the features are based on pixels or patches, but not on the human body. For instance, [1] and [2] involve computation of optical flow, whose estimation may be limited

[1] CMU graphics lab: Carnegie Mellon University Motion Capture Database, http://mocap.cs.cmu.edu (2012).

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 333–340, 2013.

by a lot of constraints (e.g. the interpretation of motion); [3] and [4] regard the video sequence as a space-time volume of intensities, gradients, or other local features. However, the method in this paper concentrates on the human body and represents the body as skeleton. Figure 1 shows the architecture of our system. Despite the fact that [5] and [6] also represent human as skeleton, the extracted features are different from the features in this paper. Instead of using joint angles, [5] uses the distribution of star skeleton as features for action recognition. In [6], only the angle between two legs is used to distinguish walk from run. The features used in this paper are vectors that consist of both static information (degrees of 9 important joint angles on the human body) and dynamic information (joint angle accelerations). Our method can not only distinguish walk from run, but also recognize jump and jumping jack. After extraction of the features, HMM is used to classify the action categories. In the training phase, each action category has an individual HMM model and the parameters of the model are estimated and optimized by the training data, in order to best describe the training sequences. While recognizing, the HMM model, which matches the test sequences best, is chosen as the recognized category.

This paper is organized as follows. In section 2, the approach of feature presentation and extraction is presented. In section 3, the observation assignment of HMM in this method is introduced. In section 4, some experimental results and discussion are reported. In section 5, conclusion and future work are outlined.

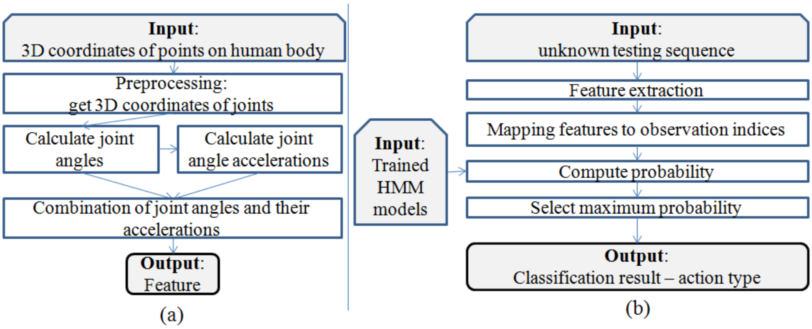

Input:
3D coordinates of points on human body
Preprocessing:
get 3D coordinates of joints
Calculate joint angles
Calculate joint angle accelerations
Combination of joint angles and their accelerations
Output:
Feature
(a)
Input:
Trained HMM models
Input:
unknown testing sequence
Feature extraction
Mapping features to observation indices
Compute probability
Select maximum probability
Output:
Classification result – action type
(b)

Fig. 1. Illustration of the system architecture. (a) Process flow of feature extraction. (b) Process flow of recognition.

2 Feature Extraction

Human action consists of a series of postures over time. The shape of human skeleton can describe the posture well. In order to present the skeleton clearly and accurately, joint angles are used. The feature we used in this paper is the vector that consists of two parts: the joint angles in radian and the joint angle accelerations.

2.1 Joint Angles

A joint is the connection location of two or more bones. Joints are very important to human, because they are constructed to allow movement and provide mechanical support. Joint angle in this paper means the angel that takes the joint position as the vertex and takes the two bones, which are connected by this joint, as the sides (as shown in Fig. 2(d)). According to the importance of body parts for describing action, 11 points on human body are chosen, namely head, neck, elbows, hands, waist, knees and feet (as shown in Fig. 2(b)). As indicated in Fig. 2(c), these 11 points construct 10 vectors. We take vector 6, i.e. the torso vector, as the base vector, and then each other vector has an angle with it. Therefore, we get 9 joint angles as shown in Fig. 2(d).

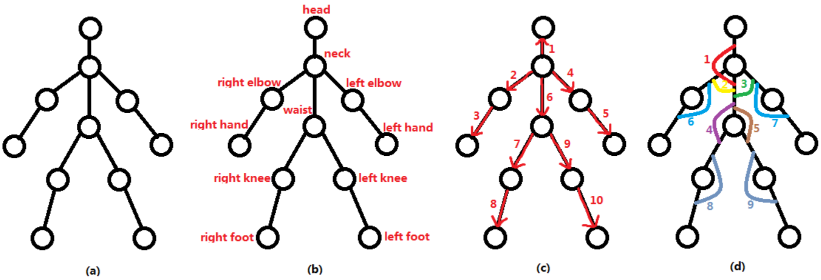


Fig. 2. Description of human skeleton. (a) Human skeleton that consists of 11 key points on the body. (b) Human skeleton with labels of the points. (c) The skeleton is described as 10 vectors; the numbers are the indices of the vectors. (d) The skeleton is described as 9 joint angles; the numbers are the indices of the angles.

2.2 Acceleration of Joint Angles

The joint angle acceleration is used as part of the feature in our method. It is useful for recognizing human action, because the amplitudes of arm swing are almost the same for some actions, e.g. walk and run, but the frequencies of the change are different, which can be described by angle acceleration. Another advantage to consider angle acceleration in the proposed method is that the dynamic features are added to HMM model, which generally only considers static states. In physics, acceleration is the rate at which the velocity of an object changes. It can be calculated by the following formula:

$$a = \frac{\Delta v}{\Delta t} \tag{1}$$

Here, we need the angle acceleration, i.e.Δv is the angle velocity and Δt is the time interval. Assume the change of angle is $\Delta\alpha$, the frame rate of the video is f fps (frame per second) and we sample the video every n frames. Then

$$\varDelta t = \frac{n}{f}, \quad \varDelta v = \frac{\varDelta \alpha}{\varDelta t} \tag{2}$$

According to formula (1) and (2), we can get

$$a = \frac{\varDelta \alpha}{\varDelta t^2} = \frac{\varDelta \alpha \cdot f^2}{n^2} \tag{3}$$

2.3 CMU Graphics Lab Motion Capture Database

The Carnegie Mellon University (CMU) motion capture database is a free dataset for use in research projects. There are totally 2605 trials in 6 categories and 23 subcategories. To capture human actions, totally 41 markers are placed on the human body as shown in Fig. 3(a). As indicated in section 2.1, we only choose 11 key points on the human body. In order to use the CMU motion capture database in our method, we use the following markers from the 41: RFHD, CLAV, LELB, LWRB, RELB, RWRB, STRN, LKNE, LANK, RKNE, RANK. The corresponding relationship is shown in Fig. 3(b). The database provides several file formats to record human actions. The file format used in this paper is c3d file. We get the 3-dimensional coordinates of the 11 markers in each frame and the frame rate from the corresponding c3d file. After that, we can calculate the joint angles according to the following formulas: (assume vector $a = (x_1, y_1, z_1)$ and vector $b = (x_2, y_2, z_2)$, α is the angle between a and b)

$$\alpha = \arccos \frac{x_1 \cdot x_2 + y_1 \cdot y_2 + z_1 \cdot z_2}{\sqrt{x_1^2 + y_1^2 + z_1^2} \cdot \sqrt{x_2^2 + y_2^2 + z_2^2}} \tag{4}$$

2.4 Invariance of the Feature

The proposed method is invariant to scale, coordinate, and transition.

Scale Invariance. As indicated in section 2.2, the features are vectors that consist of the 9 joint angles in radian and their accelerations. As the principle of angle said, no matter how long the sides are, the angle stays constant. Therefore, no matter how big or how small the human skeleton is, the joint angles do not change.

Coordinate System and Transition Invariance. According to formula (4), vector coordinates are used for calculating the joint angle. Assume there are two points, $RFHD(a_1, b_1, c_1)$ and $CLAV(a_2, b_2, c_2)$, whose positions are shown in Fig. 3(b). Then vector 1, which is shown in Fig. 2(c), is $\boldsymbol{V} = (a_1 - a_2, b_1 - b_2, c_1 - c_2)$, i.e. only the relative position of the two points is needed to calculate the vector, i.e. the feature is independent to which coordinate system is chosen. Accordingly, no matter the human skeleton appears in which corner of the video, or through which direction the person moves, as long as he is still in the sight of camera and does the same action, the skeleton shape will almost keep the same.

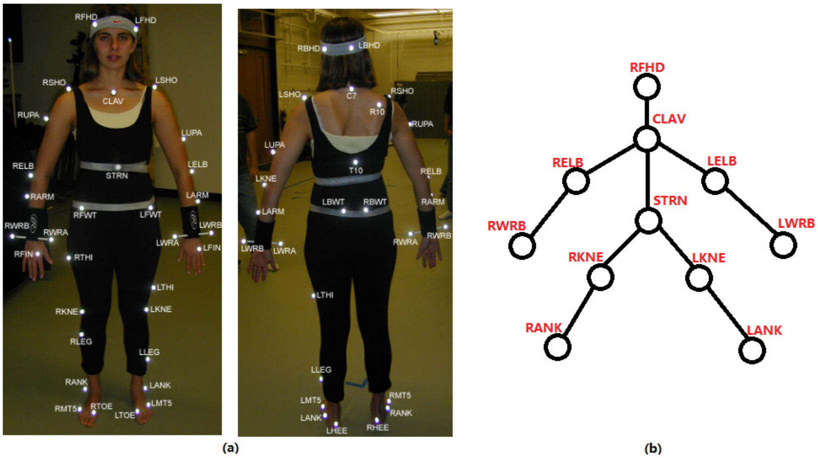


Fig. 3. Marker Illustration. (a) Marker Placement on Human Body of CMU Motion Capture Database. (b) Corresponding Relationship between Markers of CMU Motion Capture Database and Skeleton used in the Proposed Method.

3 Observation Assignment of Hidden Markov Model

HMM is a very famous and widely used model in machine learning, especially for recognition. In this paper, we use HMM as the classification tool. The word hidden in HMM means that the state of the Markov chain is actually invisible. We can only see the observations and use a transition matrix to find relationship between states and observations. Figure 4 shows a simple first-order HMM, where z_t represents state and x_t represents observation. In the proposed method, the shapes of human skeleton are used as observations, and the features - vector of joint angles and their accelerations - are used to describe human skeleton. While using HMM, we need to give each observation an index, i.e. each possible combination of feature elements will have an individual id. In this paper, we assign the index according to the following formula:

$$\boldsymbol{Index}_{Observation} = W\left(N\right) + \sum_{n=1}^{N-1} M^{n}\left(W\left(N-n\right)-1\right) \tag{5}$$

N : sum of joint angles' number and accelerations' number
M : number of intervals
W(n) : weight value of the nth joint angle or acceleration, n is the angle's index as shown in Fig. 2(d)

As indicated in formula (4), the angle is calculated by *arccos*, i.e. the angle is between 0 and π. In order to properly describe the skeleton shape, we divide interval $[0, \pi]$ into several sub-intervals to avoid too precise results and get the computational cost lower. Similarly, the acceleration is in the range of $[-\infty, +\infty]$

and should be divided into the same number of intervals as joint angles. Each interval has a weight, which is just equal to the index of interval. For example, assume that we have totally 9 angles, 9 accelerations and divide [0, π] into 3 intervals, namely the first interval [0, $\pi/3$), the second interval [$\pi/3$, $2\pi/3$) and the third interval ($2\pi/3$, π]. Similarly, we also divide [-∞, +∞] into 3 intervals. Then x is equal to 18, y is equal to 3 and assume angle 3 in Fig. 2(d) is equal to 45 degree, then *W(3)* is equal to 1 (because 45 degree is in the first interval [0, $\pi/3$), so the weight is 1).

4 Experimental Results

To evaluate the performance of our method, we implement a system to recognize totally 4 types of human actions: walk, run, jump and jumping jack, in both CMU motion capture database and Weizmann classification database[7]. Both experiments use CMU motion capture database as the training dataset. In order to train HMM models, the CMU motion capture database is split into two sets randomly: the training set and the testing set. Each HMM model is trained with 30 iterations. The illustration of skeleton motions for the 4 actions is shown in Fig. 5. According to formula (5), if x joint angles and y intervals are used, then we will have totally M^N observations. Due to the limitation of memory size, we simplify the test case, i.e. for both training and testing, 6 joint angles and 3 intervals are used.

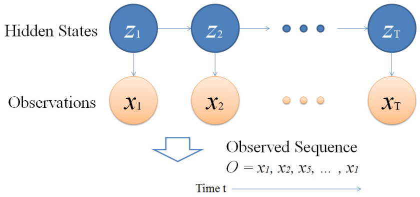


Fig. 4. A first-order HMM

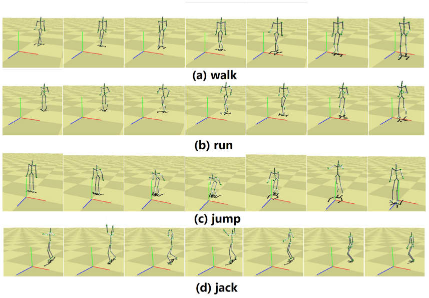


Fig. 5. Illustration of Human Skeleton Motions for the four actions

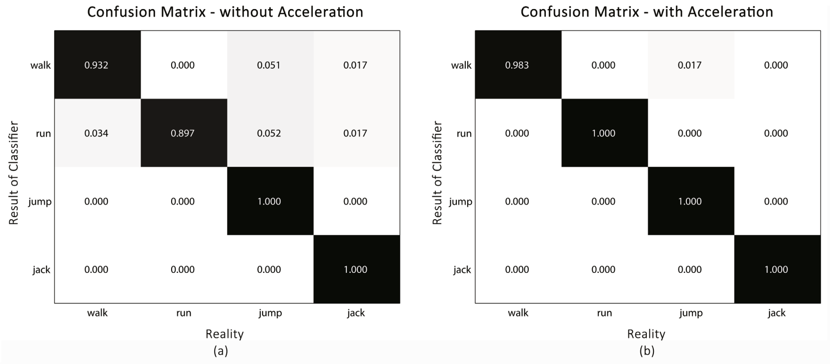
4.1 Recognition Results on CMU Motion Capture Database

We take a walk video as an example to explain the recognition process. After feature extraction and mapping features to observation indices, we get an observation sequence O = (8, 8, 2, 1, 3, 3, 6, 6, 4, 4). Then we use each trained action HMM model to compute the probability of generating sequence O, and the log scale of probabilities are as following:

```
Log(P(O|jack)) = -Inf
Log(P(O|jump)) = -261.6262
Log(P(O|run)) = -340.7212
Log(P(O|walk)) = -5.8181
```

The HMM model of walk has the maximum probability, so the video is recognized as walk.

In order to show the effect of considering angle acceleration, we compare the recognition results between method considering acceleration and method without considering acceleration. Both methods use the same frame (6 angles, 3 intervals and HMM as classification tool), the only difference is that the method considering acceleration additionally adds two dimensions to the feature – the accelerations of shoulders and thighs. The confusion matrices of recognition results are shown in Fig. 6. As indicated in Fig. 6, the method considering accelerations of joint angles get better results than the one without considering accelerations, especially for walk (a promotion of 5.1%) and run (a promotion of 10.3%).



Confusion Matrix - without Acceleration

Result of Classifier \ Reality	walk	run	jump	jack
walk	0.932	0.000	0.051	0.017
run	0.034	0.897	0.052	0.017
jump	0.000	0.000	1.000	0.000
jack	0.000	0.000	0.000	1.000

(a)

Confusion Matrix - with Acceleration

Result of Classifier \ Reality	walk	run	jump	jack
walk	0.983	0.000	0.017	0.000
run	0.000	1.000	0.000	0.000
jump	0.000	0.000	1.000	0.000
jack	0.000	0.000	0.000	1.000

(b)

Fig. 6. Confusion Matrix of Recognition Results with and without Accelerations. (a) Results using features without considering angle acceleration. (b) Results using features with considering angle acceleration.

4.2 Recognition Results on Weizmann Database

Weizmann classification database contains 11 categories of human action and each category has 9 - 10 videos. Since the form of jump of CMU motion capture database is different from the one of Weizmann classification database, we only test three types of human action to evaluate the proposed method: walk, run and jumping jack. To get the skeletons, joints and corresponding coordinates of the joints from the AVI videos, we use another project of our lab combined with some manual estimations. The recognition results for the three actions are all 100%.

5 Conclusion

In this paper, an HMM based human action recognition method using joint angle acceleration is proposed. Experimental results show that our approach gives excellent classification performances, especially for run, jump and jumping jacks. A maximum promotion of 10.3% is obtained compared with the method without considering acceleration. However, the method has a constraint – the coordinates of the joints. The excellent results are based on accurate coordinates. Therefore, a more robust method to detect and track joints on the human skeleton is necessary in future work.

Acknowledgement. The work was supported by the National Natural Science Foundation of China (Grant No. 61272251).

References

[1] Black, M.J.: Explaining Optical Flow Events with Parameterized Spatio-Temporal Models. In: Computer Vision and Pattern Recognition, vol. 1, pp. 1326–1332 (1999)

[2] Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing Action at a Distance. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 726–733. IEEE (2003)

[3] Chomat, O., Martin, J., Crowley, J.L.: A Probabilistic Sensor for the Perception and the Recognition of Activities. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, pp. 487–503. Springer, Heidelberg (2000)

[4] Zelnik-Manor, L., Irani, M.: Event-Based Analysis of Video. In: Computer Vision and Pattern Recognition, pp. 123–130 (2001)

[5] Chen, H.S., Chen, H.T., Chen, Y.W., et al.: Human action recognition using star skeleton. In: Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks, pp. 171–178. ACM (2006)

[6] Fujiyoshi, H., Lipton, A.J.: Real-Time Human Motion Analysis by Image Skeletonization. In: Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision, pp. 15–21 (1998)

[7] Blank, M., Gorelick, L., Shechtman, E., et al.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1395–1402. IEEE (2005)

Level Set Initialization Based on Modified Fuzzy C Means Thresholding for Automated Segmentation of Skin Lesions

Ammara Masood, Adel Ali Al-Jumaily, and Yashar Maali

School of Electrical, Mechanical and Mechatronic Engineering, University of Technology, Sydney, Australia
ammara.masood@student.uts.edu.au, Adel.Al-Jumaily@uts.edu.au, Yashar.Maali@student.uts.edu.au

Abstract. Segmentation of skin lesion is an important step in the overall automated diagnostic systems used for early detection of skin cancer. Skin lesions can have various different forms which makes segmentation a difficult and complex task. Different methods are present in literature for improving results for skin lesion segmentation. Each method has some pros and cons and it is observed that none of them can be regarded as a generalized method working for all types of skin lesions. The paper proposes an algorithm that combines the advantages of clustering, thresholding and active contour methods currently being used independently for segmentation purposes. A modified algorithm for thresholding based on fusion of Fuzzy C mean clustering and histogram thresholding is applied to initialize level set automatically and also for estimating controlling parameters for level set evolution. The performance of level set segmentation is subject to appropriate initialization, so the proposed algorithm is being compared with some other state-of-the-art initialization methods. The work has been tested on clinical database of 270 images. Parameters for performance evaluation are presented in detail. Increased true detection rate and reduced false positive and false negative errors confirm the effectiveness of the proposed method for skin cancer detection.

Keywords: Skin cancer, Segmentation, Diagnosis, Thresholding, Fuzzy, Active contours.

1 Introduction

The incidence of skin cancer is rapidly increasing throughout the world and malignant melanoma is the most deadly form of skin cancer [1]. An estimated 76,250 new cases of invasive melanoma were diagnosed in the US in 2012, with an estimated 9,180 resulting in death [2]. Early diagnosis and excision can make the situation better as melanoma can be cured with a simple excision if detected earlier.

Analysis of dermoscopic images is a commonly used method for diagnosis of skin cancer but, this technique demands great deal of experience [3]. Knowledge based computerized diagnostic system can be used as a standalone warning tools for helping

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 341–351, 2013.

the physicians in early diagnosis and to provide quantitative information about the lesion for experts to be considered during biopsy decision making.

Segmentation is one of the most difficult tasks in computerized analysis of skin lesions because of the great variety of lesion shapes, colors, textures, smooth transition between the lesion and the skin, ill illumination and artifacts such as skin texture, air bubbles and hair.

Several segmentation algorithms have been proposed which can be broadly classified as 1) discontinuity based segmentation 2) Similarity based segmentation, like thresholding [4], clustering [5], and region based approach [6]. The former often depends on intensity gradients, while the latter takes advantage of pixel intensities. Some articles comparing segmentation techniques are available in literature [7, 8].

Active contour is a popular approach used to estimate boundaries in medical images. It can be categorized as: 1) parametric active contours [9] which adapt a deformable curve until it fits the object boundary. 2) Geometric active contours based on level set theory. Some of the active contour models need user intervention for initialization. Thus automatic approaches like gradient vector flow algorithm [10] and robust algorithms like adaptive snakes [11] are taking important place in literature.

Level set (LS) methods have shown effective results for segmentation of medical images. However, intensive computational requirements and regulation of controlling parameters make it a complex and time consuming method. To overcome such shortcomings, few level set initialization methods like spatial fuzzy clustering and iterative thresholding are presented in [12, 13] respectively.

Fuzzy C mean (FCM) clustering [14] and iterative histogram based thresholding [15] have also been applied individually for segmentation of dermoscopic images[16, 17]. Histogram based thresholding can be a good approach for images with well separated modes in the histogram. On the other hand FCM clustering for gray level images allows categorizing the pixels in more than two classes based on intensity values.

A new algorithm that unifies advantages of fuzzy clustering and hard thresholding schemes is presented in this paper. 3-class FCM clustering integrated with histogram analysis is used for performing the thresholding operation for obtaining a binary image, which is consequently being used for initialization of the LS and calculation of the controlling parameters for efficient curve evolution. Experimental observations showed that the proposed method is more accurate as compared to standard level set method [18], region based active contours [19, 20], spatial fuzzy clustering [12], adaptive thresholding [8] and K-mean clustering [5] based Level set segmentation.

The paper is organized as follows: Section 2 provides details of the proposed segmentation method. Section 3 discusses experimental results and Section 4 provides performance evaluation. Concluding remarks are given in Section 5.

2 Methodology

This section describes our technique for segmentation of skin cancer images. The main parts of our proposed algorithm are:

2.1 Image Pre-processing

There are certain extraneous artifacts such as skin texture, air bubbles, dermoscopic gel and hair that make border detection difficult. For reducing the effect of these artifacts the images need to be pre-processed. We found that the best segmentation results were obtained using median filter with 7x7 mask to smooth the images before segmentation. Thus, the first step of the overall process is to get a filtered image.

2.2 Fuzzy C-Mean Thresholding

FCM clustering is used to partition N objects into C classes. In this case N is equal to the number of pixels in the image i.e. $N=N_x$ x N_y and C=3 for 3-class FCM clustering. The FCM algorithm uses iterative optimization of an objective function based on a weighted similarity measure between the pixels in the image and each of the c-cluster centers. A local extremum of the objective function indicates an optimal clustering of the input data. The objective function that is minimized is given by (1)

$$Q = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij})^m \left\| z_j - v_i \right\|^2 \tag{1}$$

where $z_j \in Z$ and $Z = \{z_1, z_2, z, \dots\dots.. z_N\}$ & $v_i \in V$ where $V = \{v_1, v_2, \dots v_C\}$, $\|*\|$ is a norm expressing the similarity between any measured data value and the cluster centre; m∈ [1, ∞] is a weighting exponent and can be any real number greater than 1. Calculations suggest that best choice of m is in the interval [1.5, 2.5], so m=2 is used here as it is widely accepted as a good choice of fuzzification parameter.

The aim of FCM algorithm is to find an optimal fuzzy c-partition by evolving fuzzy partition matrix U= $[u_{ij}]$ iteratively and computing cluster centres. In order to achieve this, algorithm tries to minimize the objective function Q (1) by iteratively updating the cluster centres and the membership functions using following equations.

$$v_i = \left(\sum_{j=1}^{N} (u_{ij})^m\right) / \left(z_j \sum_{j=1}^{N} (u_{ij})^m\right) \tag{2}$$

$$u_{ij} = 1 / \left(\sum_{k=1}^{C} \left(\frac{\|v_i - u_j\|}{\|v_i - u_k\|}\right)^{\frac{2}{m-1}}\right) \tag{3}$$

The fuzzy c-partition of given data set is the fuzzy partition matrix U= $[u_{ij}]$ with i=1, 2….C and j=1, 2, 3…N, where u_{ij} indicate the degree of membership of j^th pixel to i^th cluster. Membership functions are subject to satisfy following conditions.

$$\sum_{i=1}^{C} u_{ij} = 1 \text{ for j=1,2,3,….N} ;\ 0 < \sum_{j=1}^{N} u_{ij} < N \text{ for i= 1,2….C};\ 0 \le u_{ij} \le 1$$

2.3 Selection of Threshold

After performing FCM clustering, finally each pixel is assigned to the cluster for which its membership value is maximum. Thus, pixels are divided into 3 classes based on their intensity value. Secondly, histogram analysis of image is performed. So, the histogram is obtained having n bins (n=256 for grayscale image), the intensity distribution is calculated in P partitions of histogram with P=C, using following formulation.

Distributionp1,2 = $\sum_{i=n/(c+1)}^{n/(\frac{C}{2})} p\_num(i,1)$, *Distributionp2,3* = $\sum_{i=n/(c/2)}^{n/(\frac{C}{3})} p\_num(i,1)$ …
Distributionp1,P = $\sum_{i=n/(c+1)}^{nC/(C+0.5)} p\_num(i,1)$

where i is the value of intensity bin and p_num is the pixel numbers for that intensity value. We took three clusters and consequently three histogram partitions had been used, as it gives best results for skin lesion segmentation. However, the mathematical formulations can be easily extended for n number of clusters for general segmentation problems. After evaluating the clusters with dominant intensity value, the location of appropriate threshold is obtained using the following algorithm.

if (Distributionp1,2 > Distributionp2,3 && Distributionp1,2 > Distributionp1,3 ..)
Threshold value (T) = (max (data (label=1)) +min (data (label=2))/2
if (Distributionp2, 3 > Distributionp1, 2 && Distributionp2,3 > Distributionp1,3 …)
Threshold value (T) = (max (data (label=2)) +min (data (label=3))/2
if (Distributionp1, 3 > Distributionp2, 3 && Distributionp1, 3 > Distributionp2,3 …)
Threshold value (T) = (max (data (label=1)) +min (data (label=3))/2

Finally the binary image is obtained on the basis of selected threshold value. This method of threshold selection takes into account the intensity distribution in the image. Therefore, it helps in obtaining optimum threshold values for different images obtained under different conditions. FCM thresholding algorithm is presented in Fig. 1. The output of this stage is a binary image (Bi) which has been used in the following steps.

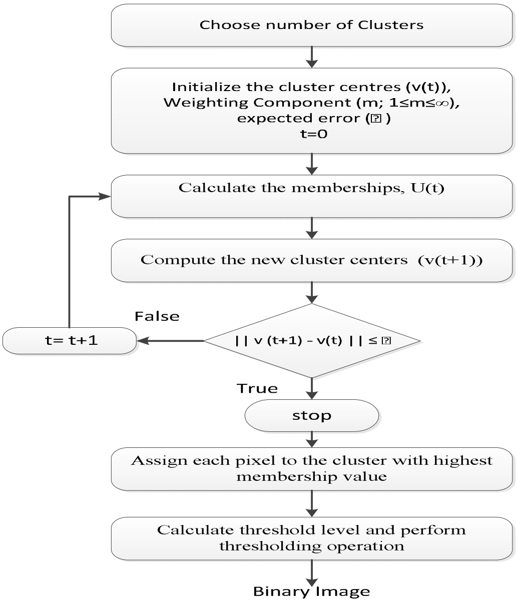

Choose number of Clusters
Initialize the cluster centres (v(t)), Weighting Component (m; 1≤m≤∞), expected error (⍰) t=0
Calculate the memberships, U(t)
Compute the new cluster centers (v(t+1))
False
t= t+1
| | v (t+1) - v(t) | | ≤ ⍰
True
stop
Assign each pixel to the cluster with highest membership value
Calculate threshold level and perform thresholding operation
Binary Image

Fig. 1. Algorithm for FCM Thresholding

2.4 Fuzzy C-mean Thresholding Based Level Set Segmentation

Active contour [19] is a well-established approach used for segmentation. It is used to detect objects using curve evolution, subject to constraints from a given image, for detecting objects. In problems of curve evolution, level set methods have been extensively used. LS methods are established on dynamic implicit interfaces and partial differential equations. In traditional LS formulation [21], the contours denoted by C, are represented by the zero level set C(t)= {(x,y)| $\emptyset$ (t,x,y)=0) of a level set function $\emptyset$ (t,x,y). The evolving equation of the LS function $\emptyset$ can be written in the following general form (4)

$$\frac{\partial\emptyset}{\partial t} + F|\nabla\emptyset| = 0 \tag{4}$$

which is called levels set equation. The function F is the speed function that represents the comprehensive forces, including the internal force from the interface geometry and the external force from image gradient or/and artificial momentums.

In order to stop the level set evolution near the optimal solution, the advancing force has to be regularized by an edge indication function g. The edge indication function used is given by (5)

$$g = 1/(1 + |\nabla I^*|^2) \tag{5}$$

where I^* is the filtered image. Traditional LS method is computationally intensive and has limitations like need of re-initialization of LS function to signed distance function for stable curve evolution [21]. Therefore, fast LS formulation [18] is used here. This method is computationally more efficient and can be implemented by using simple finite difference scheme. The overall iterative process for levels set evolution is given by (6).

$$\emptyset^{j+1} = \emptyset^j + \tau[\xi(g, \emptyset^j) + \mu\xi(\emptyset^j)] \tag{6}$$

where $\xi(g, \emptyset) = \delta_\varepsilon(\emptyset) div\left(g\frac{\nabla\emptyset}{|\nabla\emptyset|}\right) + g\delta_\varepsilon(\emptyset)$ is the term for attracting $\emptyset$ towards the variational boundary and $\xi(\emptyset) = \left(\nabla^2\,\emptyset - \frac{\nabla\emptyset}{|\nabla\emptyset|}\right)$ is the penalty term that forces $\emptyset$to approach the genuine signed distance function automatically.

The fast LS formulation provides a benefit of flexible initialization where roughly obtained region from thresholding can be used to construct initial LS function. Taking advantage of this facility binary image (Bi) obtained from FCM thresholding algorithm is used here for automatic initialization of the LS function $\emptyset$.

$$\emptyset_0(x, y) = -4\varepsilon(0.5 - Bi) \tag{7}$$

where ε is the regulator for dirac function [21] defined as follows:

$$\delta_\varepsilon(x) = \begin{cases} 0, & |x| > \varepsilon \\ \frac{1}{2\varepsilon}\left[1 + \cos\left(\frac{\pi x}{\varepsilon}\right)\right], & |x| \le \varepsilon \end{cases} \tag{8}$$

Controlling parameters are also adaptively determined from the binary image for regularizing the LS evolution process automatically. The weighting coefficient μ of penalty term $\xi(\emptyset)$ is taken as the ratio of area of on pixels in the binary image to its perimeter pixels. The time step τ is taken as 0.2/ μ so that $(\tau x \mu)$ remains smaller than 0.25 which is necessary to ensure stable evolution[18]. is coefficient of contour length for smoothness regulation and its value is taken as $0.1/\mu$. The value of can be increased for accelerating the evolution process but it leads to smoother contours. Thus, care must be taken especially for skin lesion images, where over smoothened images may lose significant details about boundary of lesion. The balloon force which determines the advancing direction and speed of the evolving curve is given as (9)

$$= -2(2 * \beta * Bi - (1 - \beta)) \tag{9}$$

where β is the modulating argument which is taken here as 0.5 through experimental analysis. Fig.2 shows a systematic diagram of the proposed method.

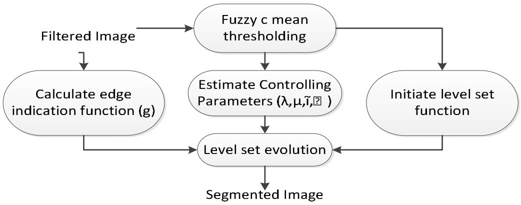

Filtered Image
Fuzzy c mean thresholding
Calculate edge indication function (g)
Estimate Controlling Parameters (λ,μ,ĩ,⍰)
Initiate level set function
Level set evolution
Segmented Image

Fig. 2. Flowchart of proposed algorithm

3 Experimental Results

Segmentation results obtained using the proposed algorithm, standard LS proposed by chan (C-LS), spatial fuzzy clustering based LS (SF-LS), Adaptive Thresholding based LS (AT-LS), and K mean clustering based LS (K-LS) are presented in Fig. 3-5 for some of the skin lesion images. While presenting the results, we tried to present images having different common problems of dermoscopic images which can badly affect the segmentation process.

It can be observed that level set method proposed initially by Chan [19] provided good segmentation in some cases but this method cannot track the border in the presence of spotty skin, many hairs or image having specular reflections as illustrated in Fig.4I(e), 4II(e), 5II(e) respectively. Similarly, K mean based LS method may result in inappropriate segmentation when the lesion color is close to skin as shown in Fig. 3I(b), 3II(b), 4I(b) Spatial fuzzy clustering and adaptive thresholding based initialization of LS improved the segmentation results in most cases but still its accuracy is less than the proposed algorithm. Analysis of the segmentation results shows that proposed method gives quite accurate results even in the presence of all these difficulties.

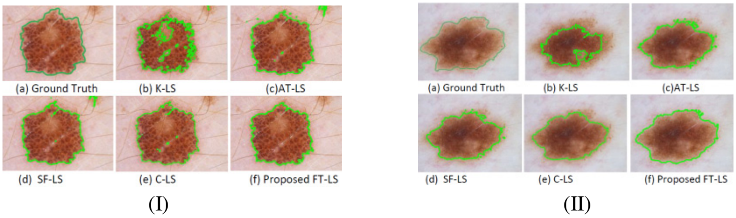


Fig. 3. Segmentation results (I) melanoma with hair and uneven border (II) dysplastic nevi with smooth transaction between skin and lesion

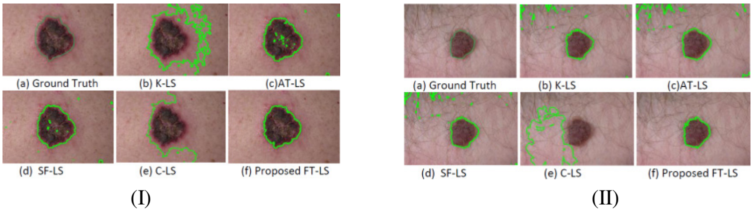


Fig. 4. Segmentation results (I) melanoma present on spotty skin with redness effect (II) benign lesion surrounded by hair

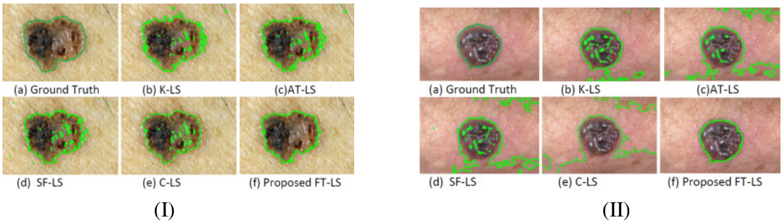


Fig. 5. Segmentation results (I) melanoma with multiple colors and prominent skin lines (II) dysplastic nevi with specular reflection and uneven illumination

4 Evaluation of Segmentation Results

In order to establish validity and clinical applicability of an algorithm, the objective evaluation of segmentation algorithms on a large set of clinical data is one of the important steps. For evaluating the efficiency of the proposed method a performance comparison is provided with some of the well-known segmentation methods used here for segmentation of same collection of skin lesion images. The metric used for measurements is based on pixel-by-pixel comparison of pixels enclosed in the segmented result (SR) and the ground truth result (GT) from expert. Mathematical details of parameters are provided below:

Hammoude Distance (HM): This metric makes a pixel by pixel comparison of the pixels enclosed by the two boundaries:

$$\mathrm{HM(SR, GT)} = \frac{\#(\mathrm{SR} \cup \mathrm{GT}) - \#\,(\mathrm{SR} \cap \mathrm{GT})}{\#(\mathrm{SR} \cup \mathrm{GT})} \tag{10}$$

It takes into account two types of error; pixels classified as lesion by automatic segmentation that were not classified as such by medical expert and pixels classified as lesion by medical expert that were not classified as such by automatic segmentation. From a clinical point of view, the 2nd type of error is more important since the lesion pixels should never be missed by diagnostic system. On the other hand, the experts demand that diagnostic system should help in reducing the rate of unwanted excision. So, the automated diagnostic system should not overestimate benign or dysplastic nevi as melanoma. Therefore, following separate metrics should be used to take into account the two types of error separately.

True Positive Rate (TPR): It measure the rate of pixels classified as lesion by both the automatic and medical expert segmentation. Higher TDR indicates better performance.

$$\mathrm{TDR(SR, GT)} = \frac{\#(\mathrm{SR} \cap \mathrm{GT})}{\#(\mathrm{GT})} \tag{11}$$

False Positive Error (FPE): It determines rate of pixels assigned as lesions by the segmentation method that were not assigned as lesion by the medical expert. Lower value of FPE indicates better performance.

$$\mathrm{FPE(SR, GT)} = \frac{\#(\mathrm{SR} \cap \mathrm{G}\bar{\mathrm{T}})}{\#(\mathrm{GT})} \tag{12}$$

False Negative Error (FNE): It determines the rate of pixels categorized as lesions by the medical expert that were not assigned as lesion by the automatic segmentation.

$$\mathrm{FNE(SR, GT)} = 1 - \frac{\#(\mathrm{SR} \cap \mathrm{GT})}{\#(\mathrm{GT})} \tag{13}$$

Dice Coefficient (DC): Dice coefficient measures agreement between the ground truth and result of automated segmentation method. It is given by the formula

$$\text{Dice Coefficient} = \frac{2\,\mathrm{TP}}{((\mathrm{FP}+\mathrm{TP})+(\mathrm{TP}+\mathrm{FN}))} \tag{14}$$

A value of 0 indicates no overlap; a value of 1 indicates perfect agreement. Higher number close to 1 indicates better agreement, and in the case of segmentation it indicates that the results match the gold standard (ground truth) better than results that produce lower Dice coefficients.

Border Error: The automatic border can be compared with manual ground truth border by calculating border error value (15), where Area (SR) represents the area inside the automatic border and Area (GT) represents the area inside the manual border.

$$\text{Border Error} = \frac{\text{Area (SR)} \cup \text{Area (GT)} - \text{Area (SR)} \cap \text{Area (GT)}}{\text{Area (GT)}} \tag{15}$$

Table 1 shows comparative segmentation results of proposed method with four other methods used for LS initialization. The database used for analysis comprised of 270 dermoscopic and clinical lesion images which were collected from various sources but

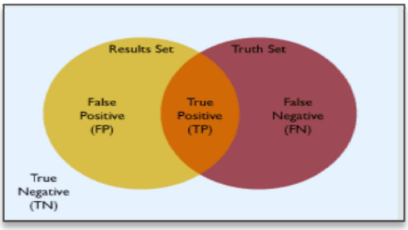


Fig. 6. Diagrammatical representation of evaluation parameters

most images were obtained from Sydney Melanoma Diagnostic Centre, Royal Prince Alfred Hospital. Segmentation results were compared with the ground truth and average of segmentation scores are presented here for each method.

Table 1. Results of Segmentation Methods. Values in bold correspond to the best performance

	Evaluated Parameters					
Method	HM (%)	TDR (%)	FPE (%)	FNE (%)	Border error (%)	DC
C-LS	23.14	89.18	11.02	10.82	13.12	0.89
K-LS	31.33	80.12	22.6	19.88	18.79	0.79
SF-LS	22.33	88.57	7.97	11.43	9.54	0.90
AT-LS	16.71	88.92	8.02	11.08	6.22	0.90
Proposed FT-LS	**11.21**	**92.86**	**4.29**	**7.14**	**4.27**	**0.94**

It is evident from the results that the proposed method has shown reasonably better performance as compared to other methods. Region based active contour method initially proposed by Chan. based on initialization using a rectangular region has shown good TDR but it has relatively high false positive error, which makes it susceptible of declaring benign lesions as melanoma. Secondly, it takes a 3 times longer time to converge to the exact border while tracking the boundary as compared to the case when the initialization process is improved. Our experimental analysis showed that K-LS gave the worst results. One reason can be because it does not take into account the fuzziness present in the skin lesion images, while differentiating between lesion and skin. Thus, initialization using K-mean results can mislead the level set method parametric contours in tracking the boundary. This shows that misleading information during initialization can lead to even worse results.

The SP-LS segmentation method using spatial fuzzy clustering for initialization and AT-LS method using adaptive thresholding for initialization reduced the false positive error and the overall border error but true detection rate of these methods is not comparable to the proposed method. On the other hand, the proposed method that is a fusion of fuzzy clustering, thresholding and LS segmentation come up with better TDR and reduced false positive, false negative and overall border error. The Dice coefficient of proposed method is quite high as compared to other methods. Thus, on the basis of our analysis, we believe that the proposed method can show promising results for lesion segmentation in a computer aided diagnosis system.

5 Conclusion

In this paper, a new segmentation algorithm is presented for skin lesion detection. It is based on the concept aimed to combine advantages of clustering, thresholding and level set methods, for getting more accurate segmentation results. The proposed approach showed good accuracy for segmentation of skin lesion images for computer aided diagnosis at the organ level. Comparative analysis proved that it works well even in the presence of different artifacts present in skin images which degrade the performance of most of the other segmentation algorithms present in literature. For researchers working in area of medical image segmentation, this method can provide basis for segmenting histo-pathological images and other types of medical images as well.

References

1. Siegel, R., et al.: Cancer statistics. CA: A Cancer Journal for Clinicians 61(4), 212–236 (2011)
2. Society, A.C., Cancer Facts & Figures (2012), http://www.cancer.org/acs/groups/content/epidemiologysurveilance/documents/document/acspc-031941.pdf
3. Piccolo, D., et al.: Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: a comparative study. British Journal of Dermatology 147, 481–486 (2002)
4. Ben Chaabane, S., et al.: Color image segmentation using automatic thresholding and the fuzzy C-means techniques. In: Proceedings of the 14th IEEE Mediterranean Electro technical Conference, pp. 857–861 (2008)
5. Dongju, L., Jian, Y.: Otsu Method and K-means. In: Proceedings of the Ninth International Conference on Hybrid Intelligent Systems, China, pp. 344–349 (2009)
6. Emre Celebi, M., et al.: Border detection in dermoscopy images using statistical region merging. Skin Research and Technology 14, 347–353 (2008)
7. Emre Celebi, M., Schaefer, G., Iyatomi, H., Stoecker, W.V.: Lesion Border Detection in Dermoscopy Images. Computerized Medical Imaging & Graphics 33, 148–153 (2009)
8. Silveira, M., et al.: Comparison of Segmentation Methods for Melanoma Diagnosis in Dermoscopy Images. IEEE Journal of Selected Topics in Signal Processing 3, 35–45 (2009)
9. Blake, A., Isard, M.: Active Contours. Springer (1998)
10. Erkol, B., Moss, R.H., Stanley, R.J., Stoecker, W.V., Hvatum, E.: Automatic lesion boundary detection in dermoscopy images using gradient vector flow snakes. Skin Research and Technology 11, 17–26 (2005)
11. Nascimento, J.C., et al.: Adaptive snakes using the EM algorithm. IEEE Transactions on Image Processing 14, 1678–1686 (2005)
12. Li, B.N., et al.: Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation. Computers in Biology and Medicine 41, 1–10 (2011)

13. Abbas, Q., Fondón, I., Rashid, M.: Unsupervised skin lesions border detection via two-dimensional image analysis. Computer Methods and Programs in Biomedicine 104, 1–15 (2011)
14. Aja-Fernandez, et al.: Soft thresholding for medical image segmentation. In: Proceedings 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Argentina, pp. 4752–4755 (2010)
15. Jun, Z., Jinglu, H.: Image Segmentation Based on 2D Otsu Method with Histogram Analysis. In: Proceedings 2008 International Conference on Computer Science and Software Engineering, pp. 105–108 (2008)
16. Tobias, O.J., Seara, R.: Image segmentation by histogram thresholding using fuzzy sets. IEEE Transactions on Image Processing 11, 1457–1465 (2002)
17. Sookpotharom, S.: Border Detection of Skin Lesion Images Based on Fuzzy C-Means Thresholding. In: Proceedings of 3rd Int. Conference on Genetic and Evolutionary Computing, China, pp. 777–780 (2009)
18. Chunming, L., et al.: Level set evolution without re-initialization: A new variational formulation. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, pp. 430–436 (2005)
19. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing 2, 266–277 (2001)
20. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic Active Contours. International Journal of Computer Vision 22, 61–79 (1997)
21. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, New York (2002)

Occlusion-Robust Face Recognition Using Iterative Stacked Denoising Autoencoder

Ying Zhang[1], Rui Liu[2], Saizheng Zhang[2], and Ming Zhu[1]

[1] Department of Automation
[2] Department of Electronic Engineering and Information Science
University of Science and Technology of China, Hefei, China
{zhy0927, liuruin, nerv1990}@mail.ustc.edu.cn, mzhu@ustc.edu.cn

Abstract. This paper investigates how to recognize faces with partial occlusions using iterative stacked denoising autoencoder (ISDAE). We introduce a mapping-autoencoder (MAE) for occlusion detection, which requires no prior knowledge of occlusion. Inspired by stacked denoising autoencoder (SDAE)'s capability to learn patterns from noisy data, we propose a novel iterative structure of SDAE for occluded faces restoration. Deep neural network (DNN) is used for final recognition. Compared with the state-of-the-art approaches (e.g. sparse representation), ISDAE achieves competitive results under serious occlusion conditions.

Keywords: face recognition, occlusion, stacked denoising autoencoder, restricted boltzmann machine, iterative, deep neural network.

1 Introduction

One challenging problem of face recognition is to recognize faces with partial occlusions. Different objects (e.g. sunglasses, scarves, extreme illumination conditions) may occlude faces, especially when face images are taken in informal ways. These occlusions can be considered as spatial noises, while the occluded faces become noisy observations of original clean faces. Various approaches have been proposed to tackle with face occlusions. Saito et al. restored the occluded areas before recognition by interpolation using principal component analysis (PCA) [1] while Hwang utilized a morphable face model [2]. Li et al. proposed a local non-negative matrix factorization (LNMF) based model [3] for partial occlusion recognition. This model learns spatially localized, parts-based subspace representation of face patterns. Wright et al. proposed a novel approach of sparse signal representation [4]. Knowing that occlusions are often sparse with regard to standard basis, they cast the problem as classifying among multiple linear regression models and this problem can be solved by sparse representation. Although this sparse coding model achieved state-of-the-art recognition results, its linear shallow structure limits its potential improvements.

In this paper, however, we exploit a deep nonlinear denoising structure addressing the occlusion problem. Recent research in deep learning indicates that deep, hierarchically learned structures perform well in difficult recognition tasks.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 352–359, 2013.

Deep network pretrained with restricted boltzmann machine (RBM) or denoising autoencoder (DAE) can capture complicated statistical patterns, and stacked denoising autoencoder (SDAE) can harness serious corruptions and distortions [5, 6, 7, 8, 13]. As we mentioned before, an occluded face can be seen as a clean face added with unknown spatial noises. Inspired by SDAE's nature of denoising and restoring, we implement it in an iterative manner for face restoration (iterative stacked denoising autoencoder (ISDAE)). To detect occluded areas, a mapping-autoencoder (MAE) is employed which requires no prior knowledge of occlusion types and positions. When "noisy" faces are restored, deep neural network (DNN) is used for robust face recognition.

In the following sections, we discuss the detail configurations of MAE, ISDAE and DNN. Furthermore, we compare our algorithm to other popular approaches to see how robust the ISDAE is for occluded face recognition.

2 Model Description

We consider an occluded face X_{occ} as a noisy version of a clean face X,

$$X_{occ} = g_{noise}(X) \tag{1}$$

In this model, $g_{noise}(\mathbf{x}) = \mathbf{x} + \mathbf{e}$, $\mathbf{e}$ is an unknown additive noise like sunglasses, scarves and masks. Our task is to first find a function $f_{\boldsymbol{\Theta}}(\cdot)$ for occluded face restoration before further recognition, where

$$\{f, \boldsymbol{\Theta}\} = \underset{f,\boldsymbol{\Theta}}{\operatorname{argmin}}\ L_H(X, f_{\boldsymbol{\Theta}}(X_{occ})) \tag{2}$$

L_H evaluates the similarity between X and the restoration $f_{\boldsymbol{\Theta}}(X_{occ})$. In our model, f is an integrated structure of MAE and ISDAE. $\boldsymbol{\Theta}$ consists of $\boldsymbol{\Theta}_{mae}$ (optimal parameters of MAE) and $\boldsymbol{\Theta}_{isdae}$ (optimal parameters of ISDAE).

Every occluded (or clean) face X_{occ} (or X) has its identity (label) y. After we get the restored face $f_{\boldsymbol{\Theta}}(X_{occ})$, we sent it to deep neural network $\phi_{\boldsymbol{\Theta}_{dnn}}(\cdot)$ for final recognition:

$$\tilde{y} = \phi_{\boldsymbol{\Theta}_{dnn}}(f_{\boldsymbol{\Theta}}(X_{occ})) \tag{3}$$

here $\boldsymbol{\Theta}_{dnn}$ is the optimal parameters of the deep neural network and $\tilde{y}$ is the output identity given by the recognition system.

2.1 Occlusion Detection with Mapping-Autoencoder

Before we use ISDAE for restoration, occlusion detection should be performed to get the positional information of occluded areas, which is critical for ISDAE to accurately restore the occluded face. Here we create a novel structure called mapping-autoencoder (MAE) for occlusion detection. MAE requires no prior knowledge of occlusion, similar strategy is adopted by [9, 10].

MAE is an autoencoder that its inputs are small patches of faces combined with patches' position maps. It is pretrained with layer-wise RBM and fine-tuned using backpropagation schemes. Given a face X_0 (X_0 can be either occluded or clean), a small fixed-size window moves across the face and densely samples several patches. For the ith patch I_b^i of X_0, we create a position map M_b^i (the same size of I_b^i) which denotes I_b^i's position in X_0,

$$M_b^i = resize(map(I_b^i, X_0), I_b^i) \tag{4}$$

$resize(\mathrm{A}, \mathrm{B})$ resets patch A's size so that A has the same size of B. $map(I, X)$ is defined as follows:

$$map(I, X)(x, y) = \begin{cases} 1, & if\ X(x, y) \in I \\ 0, & if\ X(x, y) \notin I \end{cases} \tag{5}$$

Mixing up I_b^i and M_b^i, we get an input of MAE, T_b^i, where $T_b^i = [I_b^i; M_b^i]$. To train the MAE, we only use training examples extracted from clean faces. After initializing $\mathbf{W}_{mae}$, $\mathbf{u}_{mae}$ and $\mathbf{z}_{mae}$ by pretraining, the MAE is fine-tuned by minimizing the cross-entropy error L_H between the output map $\tilde{M}_b$ and the input map M_b, see Fig.1.

To judge whether a patch I_{occ} is occluded in an occluded face X_{occ}, we evaluate the difference between I_{occ}'s output map $\tilde{M}_{occ}$ and its input map M_{occ}. MAE learns statistical patterns only from clean training examples, if I_{occ} is extracted from occluded areas, $\tilde{M}_{occ}$ will be very different from M_{occ}. Thus for I_{occ}, its probability to be occluded can be modeled using both the spatial distance $dist(\cdot, \cdot)$ and cross-entropy $L_H(\cdot|\cdot)$ between $\tilde{M}_{occ}$ and M_{occ},

$$p(occluded|I_{occ}) = \frac{2}{exp(-(\alpha L_H(\tilde{M}_{occ}, M_{occ}) + \beta dist(\tilde{M}_{occ}, M_{occ}))) + 1} - 1 \tag{6}$$

α and β are positive coefficients of $L_H(\cdot|\cdot)$ and $dist(\cdot, \cdot)$. This formula suggests that the larger the $L_H(\tilde{M}_{occ}, M_{occ})$ or the $dist(\tilde{M}_{occ}, M_{occ})$ is, the more possible that the I_{occ} is occluded. At this time, for a trained MAE $f_{\mathbf{\Theta}_{mae}}$ we have $\mathbf{\Theta}_{mae} = \{\mathbf{W}_{mae}, \mathbf{u}_{mae}, \mathbf{z}_{mae}, \alpha, \beta\}$. If there is an occluded face X_{occ}, the MAE will move through all the possible positions of X_{occ}, calculate the probability $p(occluded)$ and assign it to each pixel of X_{occ}. Finally, we get an occlusion-probability map P_{occ} that indicates the possible occlusion distribution of X_{occ}.

As the key part of pretraining, RBM is an undirected two-layer model with hidden layer $\mathbf{h}$ and input layer $\mathbf{x}$. The symmetric connections between $\mathbf{h}$ and $\mathbf{x}$ is described by weights $\mathbf{W}$ and biases $\mathbf{u}$, $\mathbf{z}$. Each layer has no inner-connections. A marginal probability of $\mathbf{x}$ in RBM is defined using an energy model,

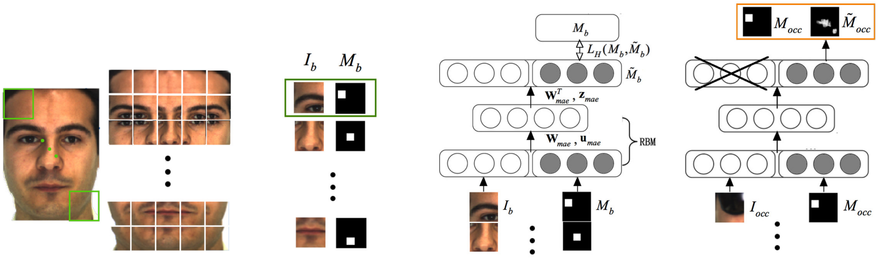
$$p(\mathbf{x}) = \sum_{\mathbf{h}} \frac{exp(\mathbf{h}^T \mathbf{W} \mathbf{x} + \mathbf{u}^T \mathbf{x} + \mathbf{z}^T \mathbf{h})}{Z} \tag{7}$$

where Z is the partition function and the conditional probabilities of $p(\mathbf{h}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{h})$ are given as follows:

$$p(\mathbf{h}_i = 1|\mathbf{x}) = sigm(\mathbf{W}_i \mathbf{x} + \mathbf{z}_i) \tag{8}$$

$$p(\mathbf{x}_j = 1|\mathbf{h}) = sigm(\mathbf{W}_j \mathbf{h} + \mathbf{u}_j) \tag{9}$$

here $sigm(\cdot)$ is the *sigmoid* function. To train a RBM, we use contrastive divergence to estimate the gradient steps [11].



(a) moving window samples small unoccluded areas as parts of training examples

(b) integrate the unoccluded area and its spatial map as the training example

(c) training occlusion detector : mapping-autoencoder pretraining with RBMs

(d) occlusion detection: compare the original map M_{occ} with the output map $\tilde{M}_{occ}$

Fig. 1. A small window moves across the face to sample patches (a). Patches are combined with their position maps to form the inputs of MAE (b). We launch a one-layer MAE pretrained by layer-wise RBM (c). Given a patch I_{occ}, its probability to be occluded is estimated based on the difference between $\tilde{M}_{occ}$ and M_{occ} (d).

2.2 Face Restoration by Iterative Stacked Denoising Autoencoder

In this section, we first give brief discussion about DAE and SDAE, then we show the whole structure of iterative stacked denoising autoencoder (ISDAE) and explain how to implement it for occluded face restoration.

As the basic building block of SDAE, DAE is a three-layer neural network that try to reconstruct the original clean input from its noisy version. Let $\mathbf{x}$ be the original input and $\tilde{\mathbf{x}}$ be the noisy version of $\mathbf{x}$ where $\tilde{\mathbf{x}} = q_{noise}(\mathbf{x})$, DAE includes the denoising encoder f_{en} and decoder f_{de},

$$\mathbf{y} = f_{en}(\tilde{\mathbf{x}}) = sigm(\mathbf{W}_{en}\tilde{\mathbf{x}} + \mathbf{b}_{en}) \tag{10}$$

$$\hat{\mathbf{x}} = f_{de}(\mathbf{y}) = sigm(\mathbf{W}_{de}\mathbf{y} + \mathbf{b}_{de}) \tag{11}$$

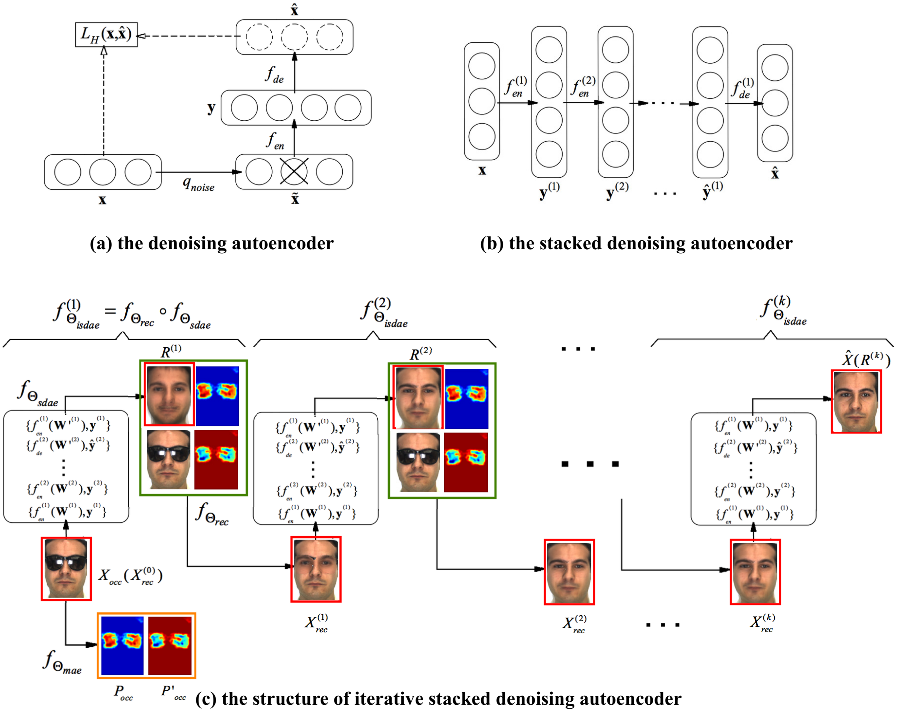
$\hat{\mathbf{x}}$ is the denoising version of $\tilde{\mathbf{x}}$, $\{\mathbf{W}_{en}, \mathbf{b}_{en}, \mathbf{W}_{de}, \mathbf{b}_{de}\}$ is weights and biases, and $\mathbf{y}$ is the encoded patterns, see Fig.2(a).

SDAE is a hierarchical structure made up of several DAEs in a stacking manner. If a SDAE consists of n DAEs and the kth DAE is made up of $f^{(k)}_{en}$, $f^{(k)}_{de}$, this SDAE can be divided into the encoding part consisting of $f^{(1)}_{en}$ to $f^{(n)}_{en}$ and the decoding part consisting of $f^{(n)}_{en}$ to $f^{(1)}_{de}$, see Fig.2(b). Considering SDAE's hierarchical characteristics, given an input $\mathbf{x}$ we have:

$$\hat{\mathbf{x}} = f^{(1)}_{de} \circ ... f^{(n)}_{de} \circ f^{(n)}_{en} ... \circ f^{(1)}_{en}(\mathbf{x}) \tag{12}$$

where $\hat{\mathbf{x}}$ is the denoising version of $\mathbf{x}$. To train these n DAEs, we adopt the similar strategy of [6].

Because of its blindness to occlusion information, SDAE performs poorly when directly used for face restoration. Therefore, we propose ISDAE, which integrates SDAE and the occlusion-probability map mentioned in section 2.1. ISDAE is a kind of iterative structure in which a restoration function $f_{\mathbf{\Theta}_{isdae}}$ is employed in



(a) the denoising autoencoder

(b) the stacked denoising autoencoder

(c) the structure of iterative stacked denoising autoencoder

Fig. 2. The basic structure of denoising autoeoncoder (a). Stacked denoising autoencoder (b). The structure of ISDAE (c). After obtaining the occlusion-probability map, occluded face is assigned to $f_{\mathbf{\Theta}_{isdae}}$ for iterative restoration.

each iteration. $f_{\mathbf{\Theta}_{isdae}}$ is made up of two parts: a SDAE $f_{\mathbf{\Theta}_{sdae}}$ and a recovering function $f_{\mathbf{\Theta}_{rec}}$. $f_{\mathbf{\Theta}_{sdae}}$ is trained using clean faces and $f_{\mathbf{\Theta}_{rec}}$ utilizes occlusion-probability map for further recovery. Suppose that there is an occluded face image X_{occ} and we already get its occlusion-probability map P_{occ} through MAE $f_{\mathbf{\Theta}_{mae}}$. X_{occ} is first sent to $f_{\mathbf{\Theta}_{sdae}}$ for preliminary restoration:

$$R = f_{\mathbf{\Theta}_{sdae}}(X_{occ}) \tag{13}$$

R is the preliminary restoration result. Occluded areas of X_{occ} are reconstructed in R using $f_{\mathbf{\Theta}_{sdae}}$'s learned statistical face patterns [6]. However, clean areas in X_{occ} are corrupted in R because they are negatively influenced by occluded areas. In this case, we select the reconstructed occluded areas of R using the occlusion-probability map, and we mix it with the clean areas of X_{occ} to get the second restoration X_{rec} in an iteration,

$$\begin{aligned} X_{rec} &= f_{\mathbf{\Theta}_{rec}}(R, X_{occ}, P_{occ}, P'_{occ}) \\ &= \langle R, h(P_{occ})\rangle^{*} + \langle X_{occ}, h(P'_{occ})\rangle^{*} \end{aligned} \tag{14}$$

$\langle \cdot, \cdot \rangle^*$ is a vector operator, $\mathbf{z} = \langle \mathbf{x}, \mathbf{y} \rangle^*$ means that $\mathbf{z}_i = \mathbf{x}_i \cdot \mathbf{y}_i$ for all $i \in \{1, ..., n\}$, n is the size of $\mathbf{z}$. $P'_{occ} = I - P_{occ}$ is a probability map of clean areas. $h(\cdot)$ is a correcting function adjusting P_{occ} or P'_{occ} in the range of $[0, 1]$ (step function is often used here). X_{rec} can also be written as follows:

$$X_{rec} = f_{\mathbf{\Theta}_{isdae}}(X_{occ}) = f_{\mathbf{\Theta}_{rec}} \circ f_{\mathbf{\Theta}_{sdae}}(X_{occ}) \tag{15}$$

A complete face restoration process includes k iterations of $f_{\mathbf{\Theta}_{isdae}}$,

$$\hat{X} = f^{(k)}_{\mathbf{\Theta}_{isdae}} ... \circ f^{(2)}_{\mathbf{\Theta}_{isdae}} \circ f^{(1)}_{\mathbf{\Theta}_{isdae}}(X_{occ}) \tag{16}$$

$\hat{X}$ is the final restoration. In the ith iteration, there are two temporary results $R^{(i)}$ and $X^{(i)}_{rec}$. $f^{(k)}_{\mathbf{\Theta}_{isdae}}$ becomes $f^{(k)}_{\mathbf{\Theta}_{sdae}}$ in the final (kth) iteration, and we have $\hat{X} = R^{(k)}$. k is the minimum number that makes the following inequation hold:

$$\varepsilon(R^{(k)}, R^{(k-1)}) < \varepsilon_0 \tag{17}$$

Here $\varepsilon(\cdot, \cdot)$ can be an error function (i.e. l^2-norm) and ε_0 is a predefined threshold.

2.3 Recognition Using Deep Neural Network

After restoration, the deep neural network (DNN) is implemented for final recognition. During DNN's training, layer-wise pretraining scheme is necessary for DNN to gain a better initialization of its multi-layer weights. In this paper, we employ SDAE $f_{\mathbf{\Theta}_{sdae}}$ in section 2.2 for DNN's initialization. We argue that pretraining using the same SDAE of restoration process can improve DNN's recognition rate.

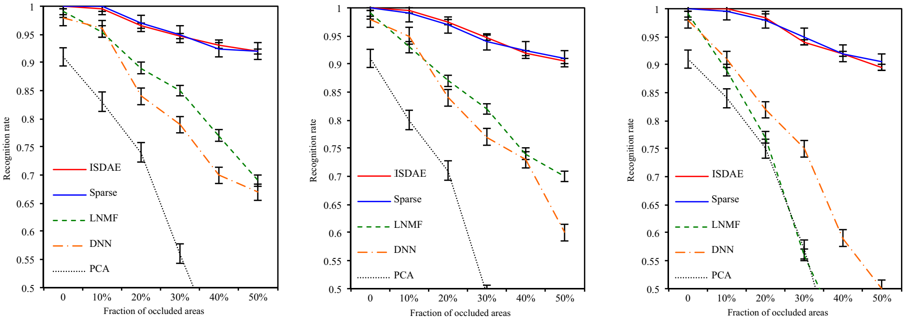
3 Experimental Results

In this section, we test our algorithm under different occlusion conditions on AR face database [12]. The AR face database contains more than 4000 face images corresponding to 126 individuals with different facial expressions, illumination conditions and occlusions (sunglasses and scarves). There are 26 pictures taken in two different sessions for each individual, and 14 of them are clean faces.

In our experiments, we randomly choose a subset of the database consisting of 40 males and 40 females. All the raw images are cropped to contain face areas only, each face is resized to 60 × 45 pixels and converted to grayscale. We compare our algorithm to PCAs (traditional, robust), LNMF, DNN (pure neural network) and sparse representation. We conduct two experiments: first we compare the selected algorithms on faces wearing sunglasses and scarves. Then we manually add different noises on clean faces and evaluate selected algorithms under various occlusion levels. The following configuration of hyperparameters (e.g. layer-parameters of MAE, SDAE and DNN) are locally optimal. In the first experiment, the training data are clean faces from all the two sessions. The detecting window's size is 12 × 12 pixels, and its moving step is 2 pixels for

Table 1. Recognition Rates using Different Algorithms

OccTypes\Methods	ISDAE	Sparse	LNMF	DNN	PCA
Sunglasses	**97.0**%	**97.0**%	65.5%	61.3%	68.8%
Scarves	**93.8**%	93.5%	50.2%	55.9%	12.0%



(a) Occlusion on the upper part (especially around eyes)

(b) Occlusion on the lower part (especially around mouth)

(b) Randomly located occlusion

Fig. 3. ISDAE is compared with sparse representation, LNMF, DNN and PCAs under occlusion conditions: occlusions on the upper part of face (around eyes) (a), occlusions on the lower part of face (around mouth) (b) and occlusions at random locations (c).

training and 1 pixel for detecting. We launch a 288-100-288 MAE. The SDAE's structure is 2700-1000-800-800. Only the first layer DAE is trained using 20% random masking noise. The noisy inputs pass on to the first layer to produce the second layer DAE's noisy input and we repeat this process in higher layers. DNN simulates the same layer structure of SDAE. The result is showed in Table 1. As we can see, ISDAE performs much better than LNMF, DNN and PCA, and it achieves slightly better recognition rates than sparse representation. In the second experiment, three kinds of occlusions are evaluated: occlusion on the upper part of face (around eyes), occlusion on the lower part of face (around mouth) and random located occlusion. The types of occlusion include black/white masks and Gaussian noises. The shapes of occlusion contain rectangles, spots and irregular shapes. We simulate various levels of occlusion: 10%, 20%, 30%, 40% and 50%. Unlike the first experiment, here we use session 1's faces for training and add noises to session 2's faces for testing. The structures of MAE, ISDAE and DNN are the same as in the first experiment. The result is showed in Fig.3. Here both the recognition rates of ISDAE and sparse representation decrease very slowly when occlusion level increases. These two methods achieve approximately the same good results under serious occlusion conditions (occlusion level $\geq$ 40%), suggesting that our method is competitive comparing to sparse representation. Other methods get poor performance when occlusion level $>$ 20%.

4 Conclusion

In this paper, we present a novel deep learning based algorithm for occluded face recognition, where MAE, ISDAE and DNN are combined for occlusion detection, restoration and recognition. Both MAE and ISDAE are new models which inherit DAE's denoising nature and are more robust to serious face occlusions than traditional approaches. These deep nonlinear denoising structures could be employed to other denoising tasks (e.g. image inpainting, object recognition under extreme illumination conditions).

Acknowledgement. This research was supported by the "Strategic Priority Research Program: Network Video Communication and Control" of the Chinese Academy of Sciences (Grant No. XDA06030900).

References

1. Saito, Y., Kenmochi, Y., Kotani, K.: Estimation of eyeglassless facial images using principal component analysis. In: IEEE ICIP (1999)
2. Hwang, B.W., Lee, S.W.: Reconstruction of partially damaged face images based on a morphable model. IEEE TPAMI 25(3), 365–372 (2003)
3. Li, S.Z., Hou, X.W., Zhang, H.J., Cheng, Q.S.: Learning spatially localized, part-based representation. In: IEEE CVPR (2001)
4. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust Face Recognition via Sparse Representation. IEEE TPAMI 31(2), 210–227 (2008)
5. Hinton, G.E.: A Practical Guide to Training Restricted Boltzmann Machines. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade, 2nd edn. LNCS, vol. 7700, pp. 599–619. Springer, Heidelberg (2012)
6. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. Journal of Machine Learning Research 11, 3371–3408 (2010)
7. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research 11, 625–660 (2010)
8. Bengio, Y.: Learning Deep Architectures for AI. Foundations and Trends in Machine Learning 2(1), 1–127 (2009)
9. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
10. Luo, P., Wang, X.G., Tang, X.O.: Hierarchical Face Parsing via Deep Learning. In: IEEE CVPR (2012)
11. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation 14, 1771–1800 (2002)
12. Martinez, A., Benavente, R.: The AR face database. CVC Tech. Report 24 (1998)
13. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)

Image Classification Based on Weight Adjustment before Feature Pooling

Shaokun Feng, Hongtao Lu, and Lei Huang

MOE-Microsoft Laboratory for Intelligent Computing and Intelligent
Systems Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, 200240, P.R. China
{superkkking,htlu,tommyflynns}@sjtu.edu.cn

Abstract. In image classification based on Bag-of-Features(BoF), the Locality-constrained Linear Coding (LLC) is a successful implementation, which is a more effective coding scheme compared with the traditional vector quantization(VQ) coding. Although, to achieve the best performance, max pooling scheme is chosen in the SPM layer, much of the spatial information is still lost during the pooling step, because all the coded descriptors are given the same importance to obtain the final representation. In this paper, we propose a new scheme that makes full use of spatial structure information to readjust their relative weights red and thus give some descriptors more chances to appear in the final feature vector more than others. Experiments of image classification on benchmark datasets show that the proposed method outperforms the LLC method.

Keywords: Image Classification, Weighting Adjustment, Feature Pooling, Weight Map.

1 Introduction

The traditional Bag-of-Features (BoF) [1] framework is one of the most widely used model in classification systems. It consists of three steps, that is, extracting SIFT [2] as local descriptors, building a semantic vocabulary, quantizing descriptors onto the codebook, and finally pooling visual words to a statistical representation for the image.

However, the BoF framework ignores the geometric relationship between features, so an extension of BoF, named Spatial Pyramid Matching (SPM) [3], was proposed afterwards. The SPM method partitions the image into $2^l \times 2^l$ sub-regions in different scales, and then calculates the histogram within each sub-region. Finally, all the histogram are concatenated to form a final vector representation of the image. It achieves impressive performance for image classification, but it requires for nonlinear Mercer kernels, such as Chi-square kernel, which will consume O(n^3) computation complexity during training. Yang et al. propose an extension of the SPM approach, named Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification(ScSPM) [4]. They

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 360–367, 2013.

use sparse codes (SC) of SIFT features to computes the spatial-pyramid representation of an image instead of the K-means vector quantization (VQ) as in the traditional SPM. Also, the max spatial pooling is taken in place of the original averaging spatial pooling method. Attractive advantage is its ability to work with linear classifiers, which reduces the training computation complexity to O(n).

Further improvement was subsequently proposed by Yu et al.[5] They observed the SC results tend to be local,i.e. nonzero coefficients are often assigned to bases nearby to the encoded data, so their method, called Nonlinear learning using Local Coordinate Coding (LCC), explicitly requires the coding to be local. Later, Wang et al. present a more local coding scheme called Locality-constrained Linear Coding (LLC) [6], which can be treated as a fast implementation of LCC. Their report shows an improvement in image classification accuracy with the classifier still being linear.

Although the LLC takes a big step forward, it still lacks the consideration of the spatial information, especially when it comes to pooling. Geometric relationship between coded features within an sub-region is ignored. We observed that some features are more important than others during pooling, especially when they lied on the edges or near edges in an image.

Based on these observations, we present a novel improvement. In our scheme, the coded features will be given different weights before going on to the pooling. Our experiments show that, after combined with calculated spatial weight, the results outperform the original LLC by a margin on various applications.

The rest of this paper is organized as follows: Section 2 briefly reviews the idea of LLC. Section 3 presents our method to construct weight maps and its application to the pooling. The experimental results are given in Section 4 and conclusion is drawn in Section 5.

2 LLC

Our work is based on LLC, so we first give a brief introduction of LLC in this section. The LLC model and our improvement are shown in Figure 1.

2.1 Feature Extraction and Codebook Training

For strong discriminative power, high dimensional local descriptors, such as SIFT descriptor are favored. SIFT descriptors are extracted from 16*16 pixel patches

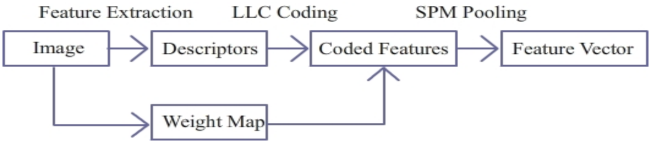


Fig. 1. The LLC modal and our modification on the modal

located by every 8 pixels. This idea is based on the comparative evaluation of Fei-Fei [7] ,whose experiments show better effect of SIFT for scene classification.

After features extracted, all feature vectors are collected to train a codebook. Traditionally, this codebook is computed by K-means algorithm. And the codebook is used for quantizing each feature vector to its corresponding codeword.

2.2 Coding Descriptors with LLC

The authors of LLC demonstrates that, locality is more essential than sparsity, so they proposed the following criteria for LLC coding:

$$\underset{C}{\text{argmax}} : \sum_{i=1}^{N} \|\mathbf{x}_i - \tilde{c}_i\|^2 + \lambda \|\mathbf{d}_i \odot \mathbf{c}_i\|^2 \tag{1}$$
$$subject\ to : \mathbf{1}^T \tilde{c}_i = 1, \forall i$$

Here, $\mathbf{x}_i$ is a local descriptor, $B = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_M]$ is the codebook entries, and $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N]$ is the set of codes for $\mathbf{X}$. By giving different weights for basis vectors, features are projected to the bases which are nearer to the feature than others. The distance is measured by:

$$\mathbf{d}_i = exp(\frac{dist(\mathbf{x}_i, \mathbf{B})}{\sigma}) \tag{2}$$

Where $dist(\mathbf{x}_i, \mathbf{B}) = [dist(\mathbf{x}_i, \mathbf{b}_i), ...dist(\mathbf{x}_i, \mathbf{b}_M)]^T$. $dist$ is the Euclidean distance, and a parameter σ is used for controlling the weight decay speed.

2.3 Feature Pooling

When all feature vectors are projected to new bases, the pooling procedure starts on the SPM layers. Coded features in a sub-region are pooled together to get a pooled vector, then these pooled vectors are concatenated and normalized, resulting in a super feature vector as the final representation of the image. Commonly used pooling strategy is the max pooling[4,6]:

$$\mathbf{c}_{out} = max(\mathbf{c}_{in1}, ..., \mathbf{c}_{in2}) \tag{3}$$

3 Weighting Adjustment

Objects in images can be simply recognized by their shapes. Their shapes offer so much discriminative information that we shouldn't discard before going on to the pooling step. The LLC applies max pooling on SPM layers, which take the image structure into account. In this paper, we will further enhance the spatial character of coded features before pooling. The adjusted features will have different importance when constructing the final feature vector. In this section, we first introduce a weight map based on the edge detection in images. Afterwards we present the weight adjustment on the coded features.

3.1 Edge Detection

Shapes in an image can be easily captured by an edge detector. There are a lot of existing edge detectors. Among them, the Canny detector achieves satisfactory effect and runs at an acceptable speed. After edge detection, the original image **I** get its response **E**, which is the edge map.

$$\mathbf{E} = Canny(\mathbf{I}) \tag{4}$$

Our experimental results show that other edge detectors give similar results. The selection of threshold plays an important role in presenting how much detail of an image. In section 4, we give a comparison of performance under different thresholds.

3.2 Weight Map

Features are densely patches sampled on a grid over an image, so some of them are near edges while others are not. Our method demonstrates that, as a feature patch goes away from its nearest edge, its importance declines at a high speed. Besides, two patches which stay the same distance to their respective nearest edges share the same importance. Based on the above observation we give the following weight map:

$$\mathbf{W} = \{\omega_{ij}\}_{H*W} \tag{5}$$

where ω_{ij} is the weight of a point in the image which is defined as:

$$\omega_{ij} = \frac{1}{2\pi\sigma^2} e^{-\frac{d_{ij}^2}{2\sigma^2}} \tag{6}$$

Here, d_{ij} is the Euclidean distance from the point to its nearest edge. This way guarantees the same distance produce same effect.

To realize the Eq.(6), we propose a fast implementation. After the edge detection, we dilate the edge map to different scales, and adjust their intensities according to their scales. Finally, sum them up to get our weight map W. The above process is depicted in Algorithm 1, and is illustrated in Figure 2:

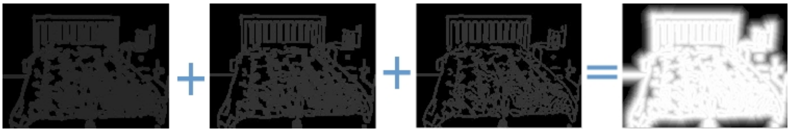


Fig. 2. The weight map is generated by the summation of dilated edge maps

3.3 Codes Adjustment

Before going on to SPM layer pooling, the coded features will be adjusted by the weight map. The new coded features becomes:

$$\mathbf{c}'_{ij} = \mathbf{c}_{ij} \times \mathbf{w}_{ij} \tag{7}$$

4 Experimental Results

In this section, we compare our method with LLC for image classification on two widely used datasets: Caltech-101[7] and 15-scene[3] dataset. They represent two important kinds of images in our life. We use a single descriptor, the SIFT, extracted from patches densely located by every 8 pixels on the image under the scale 16×16, just as LLC did. Other parameters are also share with the LLC , such as using 4×4, 2×2, and 1×1 sub-regions for SPM layer pooling and no larger than 300*300 pixels with preserved aspect ratio.We take $\sigma = 12$, lower threshold=0.04 for Caltch101 and $\sigma = 20$, lower threshold=0.004 for 15-scene.

Algorithm 1. A fast implementation of Eq.6

```
Input:
    The edge map, $\mathbf{E}$;
    The scale in gaussian function, $\sigma$;
Output:
    The weight map, $\mathbf{W}$;
Compute the scale of the largest dilated image to expand, $scale = \lfloor \sqrt{2}\sigma \rfloor$
Compute the values of gaussian function $g = \frac{1}{2\pi\sigma^2} e^{-\frac{d^2}{2\sigma^2}}$ at $d = 1, 2, ..., scale$,
    $\mathbf{g} = [g(1), g(2)...g(scale)]$
Filter $\mathbf{g}$ with [-1 1],resulting $\mathbf{b} = [b_1, b_2, ...b_{scale}], b_i = g(i) - g(i+1), g(scale+1) = 0$
for $i = 1$; $i <= scale$; $i++$ do
    Dilate the $\mathbf{W}$ by $i$ pixels, $W_i$;
end for
Sum $W_i$ up, $\mathbf{W} = \sum_{i=1}^{scale} b_i W_i$
return $\mathbf{W}$;
```

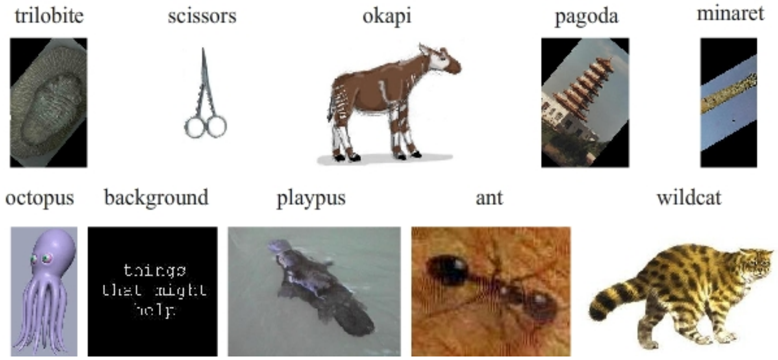


Fig. 3. Sample images from Caltech101. The accuracies of upper ones is high, while that of the lower ones are relatively low.

Table 1. Classification Accuracy on Caltech101

training images	5	10	15	20	25	30
SPM[3]	-	-	56.4	-	-	64.6
ScSPM[4]	-	-	67.0	-	-	73.2
LLC[6]	51.15	59.77	65.43	67.74	70.16	73.44
Ours	**51.74**	**61.73**	**66.65**	**69.72**	**72.03**	**73.90**

4.1 Results on Caltech101

The Caltech101[7] dataset contains 9144 images in 102 categories which includes 101 classes and a background class. The number of images per category varies from 31 to 800. Besides, deformation is applied to objects in the same category. Some sample images are shown in Figure 3. In accordance with LLC, the whole dataset is partitioned into two parts, 5-30 being training images and the others being testing images per class. The codebook we trained has 2048 bases to keep consistent with LLC as well. We compared our results with LLC in Table 1, and our method outperformed LLC by a average margin of about 1%.

4.2 Results on 15-scene

The 15-scene[3] dataset contains 4485 images of 15 scenes. It has widely been used for scene understanding task, Figure 4 shows some samples of different categories. The classification results are shown in Table 2.

4.3 Disccusion

To provide more comprehensive analysis of our proposed method, we compared the classification accuracies with different parameter settings, including different thresholds for edge detector and different values of σ in Eq.6. As can be observed

Fig. 4. Sample images from 15-scene. The accuracies of upper ones is high, while that of the lower ones is relatively low.

Table 2. Classification Accuracy on 15-scene

training images	10	20	30	50	100
SPM[3]	-	-	-	-	81.4
LLC[6]	67.45	72.48	75.57	78.41	81.96
Ours	**69.21**	**73.11**	**76.39**	**79.05**	**83.01**

in Figure 5, different parameter settings will significantly affect the resulting accuracy. Generally, higher threshold and lower σ is suitable for images in which objects is easily separated from their background, while lower threshold and higher σ is suitable for others, such as natural scenes.

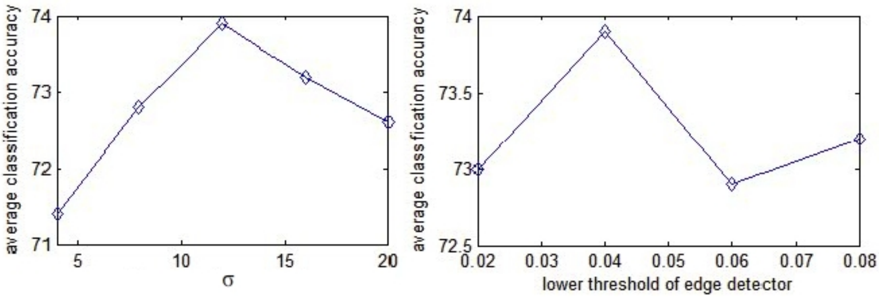


Fig. 5. Performance on Caltech 101 under different parameter settings

5 Conclusion

In this paper, we propose a novel modification on coded features before pooling for image classification. By adjustment according to weight map, the coded features will carry more spatial information. We also give an fast implementation for the weight map generation. The future work will include more spatial information in images and their combination with other classification models.

References

1. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, vol. 1, p. 22 (2004)
2. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
3. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
4. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1794–1801. IEEE (2009)
5. Yu, K., Zhang, T., Gong, Y.: Nonlinear learning using local coordinate coding. Advances in Neural Information Processing Systems 22, 2223–2231 (2009)
6. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3360–3367. IEEE (2010)
7. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 524–531. IEEE (2005)

Fast Approximated Discriminative Common Vectors Using Rank-One SVD Updates

Francesc J. Ferri[1,⋆], Katerine Diaz-Chito[1,2], and Wladimiro Diaz-Villanueva[1]

[1] Departament d'Informàtica, Universitat de València, Spain
[2] Centre de Visió per Computador, Universitat Autònoma de Barcelona, Spain

Abstract. An efficient incremental approach to the discriminative common vector (DCV) method for dimensionality reduction and classification is presented. The proposal consists of a rank-one update along with an adaptive restriction on the rank of the null space which leads to an approximate but convenient solution. The algorithm can be implemented very efficiently in terms of matrix operations and space complexity, which enables its use in large-scale dynamic application domains. Deep comparative experimentation using publicly available high dimensional image datasets has been carried out in order to properly assess the proposed algorithm against several recent incremental formulations.

1 Introduction

Dimensionality Reduction (DR) and feature extraction has always been an issue of key importance in Machine Learning and Pattern Recognition. Using appropriately reduced feature subspaces may lead to very efficient implementations of learning systems able to cope with modern challenges that involve huge amounts of data usually with very high dimensionalities. Many different application domains such as (hyper)text mining, image retrieval, stream processing or data analysis in bioinformatics will potentially require these techniques in order to achieve their corresponding goals. Supervised DR algorithms pursue the maximization of the separability among categories of objects. Linear Discriminant Analysis (LDA) is a very well-known DR approach that is based on both maximizing class separability while minimizing intraclass variability by using a simple linear transformation of the original problem [1]. For undersampled problems with more dimensions than samples, also referred to as the small sample size case (SSS), many different alternatives have been proposed to cope with the intrinsic weaknesses of LDA. Among them we can name PCA+LDA [2, 3], direct LDA [4], Null space LDA [5],or least squares LDA [6].

When really big data (or infinite streams) need to be processed, the so-called incremental algorithms are particularly appealing. In short, incremental algorithms (which can exactly reproduce the effect of a given batch algorithm or not) process a very small amount of new data to update a convenient global

[⋆] Work partially funded by FEDER and Spanish Government through projects TIN2009-14205-C04-03 and Consolider Ingenio 2010 CSD07-00018.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 368–375, 2013.

model in a very efficient way. In particular, several different incremental formulations of subspace learning algorithms have already been proposed such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) or their variants and extensions [7–12].

Subspace learning methods based on Discriminative Common Vectors (DCV) that constitute an alternative reformulation of Null space LDA [5]are particularly appealing because their good performance behavior and flexibility of implementation specially in the case of very large dimensionalities as in image recognition or genomic problems. This has motivated a recent interest in obtaining efficient implementations [13] including incremental formulations [14, 15].

In this paper, an novel incremental formulation that approximates the DCV method is proposed and evaluated on publicly available data. The algorithm consists of a per sample correction along with an additional restriction on the growth of the range space of the corresponding scatter matrix. Both subspace projection and explicit discriminative vectors can be efficiently recomputed allowing the application of these algorithms in interactive and dynamic large-scale problems as with previous incremental algorithms with some additional benefits regarding space and time complexity and generalization ability.

2 Null Space LDA and Discriminant Common Vectors

The DCV method was initially proposed for face recognition problems [5]. In particular, the method looks for a linear projection that maximizes class separability using a modified Fisher criterion that uses the within-class scatter matrix, S^w. To this end, it constructs a linear mapping onto the null space of the within-class scatter matrix in which all training data gets collapsed into the so-called *discriminant common vectors* (DCV). Classification of new data can be then done by first projecting and then measuring similarity to DCVs of each class.

Let $\mathcal{X} \in \mathbb{R}^{d\times M}$ be a given training set consisting of M d-dimensional (column) data vectors, $x_j^i \in \mathbb{R}^d$, where $i = 1, \ldots, M_j$ refers to objects of any of the c given classes, $j = 1, \ldots, c$ and $M = \sum_{j=1}^{c} M_j$, and let S_X^w be their corresponding within-class scatter matrix which can be defined as $\mathcal{X}_c\mathcal{X}_c^T$, where $\mathcal{X}_c$ is the centered version of $\mathcal{X}$, (each object in $\mathcal{X}$ with its own class mean subtracted).

Let $U \in \mathbb{R}^{d\times r}$ be a matrix whose columns are eigenvectors of S_X^w corresponding to non zero eigenvalues. r and $n = d - r$ are the dimensions of the range and null spaces of S_X^w, respectively. The j-th class common vector (CV) can be obtained as the orthogonal projection of any vector from the same class onto this null space, which can be implicitly obtained as

$$x_{com}^j = x_j^1 - UU^T x_j^1. \tag{1}$$

The representative vector x_j^1 can be substituted by any other vector in the same class or even the mean, $\overline{x_j}$, which leads to the same result [5].

Note that the method refers to the (usually huge) null space by explicitly managing its orthogonal complement, the range space of dimensionality r.

From the small set of c CVs obtained with Eq. 1, a linear mapping, $W \in \mathbb{R}^{d\times(c-1)}$, that maximizes CV scatter can be obtained using PCA or other equivalent alternatives. This mapping is the main result of the method and it serves to project both training and test data to a reduced space where a distance-based classifier is usually applied.

Instead of using centering (subtracting class means) and eigendecompositions, it is possible to use differencing (subtracting an arbitrary vector) and orthonormalization. Let $\mathcal{B} \in \mathbb{R}^{d\times(M-c)}$ be a matrix whose columns are given by difference vectors $x_j^i - x_j^1$, where $j = 1,..,c$, and $i = 2,..,M_j$. It can be shown that the range subspace of S_X^w and the subspace spanned by $\mathcal{B}$ are the same [5]. Let Q be a basis for $\mathcal{B}$ obtained through orthonormalization. It holds that $UU^T = QQ^T$ even though U and Q are different in general. Also, one can define difference CVs, $\mathcal{B}^{com}$ and do the same to obtain a mapping W equivalent to the one obtained through eigendecomposition up to a rotation.

Both strategies can be efficiently implemented using Singular Value (SVD) or QR decomposition (QRD) in their thin or economic versions [16] as it is schematized in the following algorithm [14]:

Algorithm 1. The (batch) DCV algorithms.

```
Input: $\mathcal{X} \in \mathbb{R}^{d\times M}$, dataset
Output: $W \in \mathbb{R}^{d\times(c-1)}$, mapping
1 Compute difference set, $\mathcal{B}$.                 / Compute centered set $\mathcal{X}_c$.
2 Compute $Q$ using thin QRD of $\mathcal{B}$.           / Compute $U$ using thin SVD of $\mathcal{X}_c$.
3 Obtain CVs as $x_{com}^j = x_j^1 - QQ^T x_j^1$.  / Obtain CVs as $x_{com}^j = x_j^1 - UU^T x_j^1$
4 Compute difference set, $\mathcal{B}^{com}$.           / Compute centered set, $\mathcal{X}_c^{com}$
5 Obtain $W$ using thin QRD of $\mathcal{B}^{com}$.     / Obtain $W$ using thin SVD of $\mathcal{X}_c^{com}$.
```

The computational cost of the batch approach is dominated in any case by the step 2 which is either $O(dM^2)$ or $O(dM^2 + M^3)$ that leads to the same asymptotic cost in the SSS case as $M \ll d$. In general, obtaining the basis, U, that corresponds to non zero singular (or eigen-)values leads to less efficient algorithms compared to other approaches based on orthonormalization that obtain arbitrarily rotated bases [13]. But even the fastest of these implementations which is still in $O(dM^2)$ becomes prohibitive for huge databases or data streams. The alternative consists of considering significantly faster algorithms that use an incremental strategy.

Assume that both Q and W have already been obtained and one needs to update them as a new sample $x \in \mathbb{R}^{d\times 1}$ from class k is made available. The corresponding incremental update can be obtained by extending Q with the new orthogonal vector

$$z = \frac{y - QQ^T y}{||y - QQ^T y||}$$

where $y = x - x_k^1$ and x_k^1 is the same vector that has been subtracted to all previous vectors in the kth class. Once Q has been updated, both DCVs and projection W can be obtained using steps 3–5 in the batch algorithm. Alternatively [14], one can directly update W by orthonormalizing $W - zz^TW$, or through a rank-one update of the corresponding QRD of $\mathcal{B}^{com}$ [17]. We will refer to any of these implementations as Incremental Discriminative Common Vectors (IDCV). Note that the cost is dominated by the matrix operations to obtain Q which are $O(dr)$ while computing W is $O(dc^2)$ in all cases.

3 Updating DCV Projections Using Incremental SVD

Even though the IDCV algorithm is very efficient in most situations, for very huge sets of linearly independent data an $O(dM)$ cost can be prohibitively high. An obvious choice to keep the running times of incremental updates at moderate values consists of restricting the growth of the dimension of the within-class scatter matrix range space, r, in a similar way as it has been proposed for batch DCV [18]. Basically, the idea is to discard directions in the range space which are less important. Consequently, to be able to restrict the range space, we need to incrementally compute both singular vectors and values of the corresponding centered data, $\mathcal{X}_c$, in a sufficiently efficient way. Basis vectors corresponding to small singular values will be candidates to be removed from U.

Let $\mathcal{X}_c = USV^T$ be the SVD of the current centered data and we are interested in obtaining the SVD of the updated dataset $[\mathcal{X} \mid x]_c$ which needs to be centered with regard to the updated mean. By using the incremental expression to update the mean and adding a new column to dataset, the following expression can be arrived at

$$[\mathcal{X} \mid x]_c = [\mathcal{X}_c \mid \mathbf{0}] \; + \; \frac{\overline{x_k} - x}{M_k + 1} \, [\mathbf{1}^T \mid -M_k]$$

which leads to a rank-one update of a previous SVD that can be done in time $O(dr + r^3)$ in the worst case [19]. $\mathbf{0}$ and $\mathbf{1}$ are column vectors of either zeros or ones of the appropriate dimension. The SVD update leads to an increase of r in 1 for linearly independent data. In the linear dependent case, the value of r is kept by discarding the vector whose singular value is zero [19].

Our proposal consists of measuring the importance of an update as $||y - UU^Ty||$ and if it is below a threshold, then the last singular vector corresponding to the smallest (non zero) singular value will be discarded. The threshold is set to zero initially and it is updated in such a way that the probability of discarding the new direction is zero at the beginning and tends to its maximum with iterations. A parameter ν has been empirically adjusted in such a way that the rank, r, is kept constant when $\nu = 0$ and grows linearly (for linearly independent data) when $\nu = 1$. For intermediate values, a sublinear growth is observed.

4 Experiments and Discussion

A number of experiments have been carried out to assess the relative benefits of the proposed algorithm with regard to IDCV. In the experiments the proposed

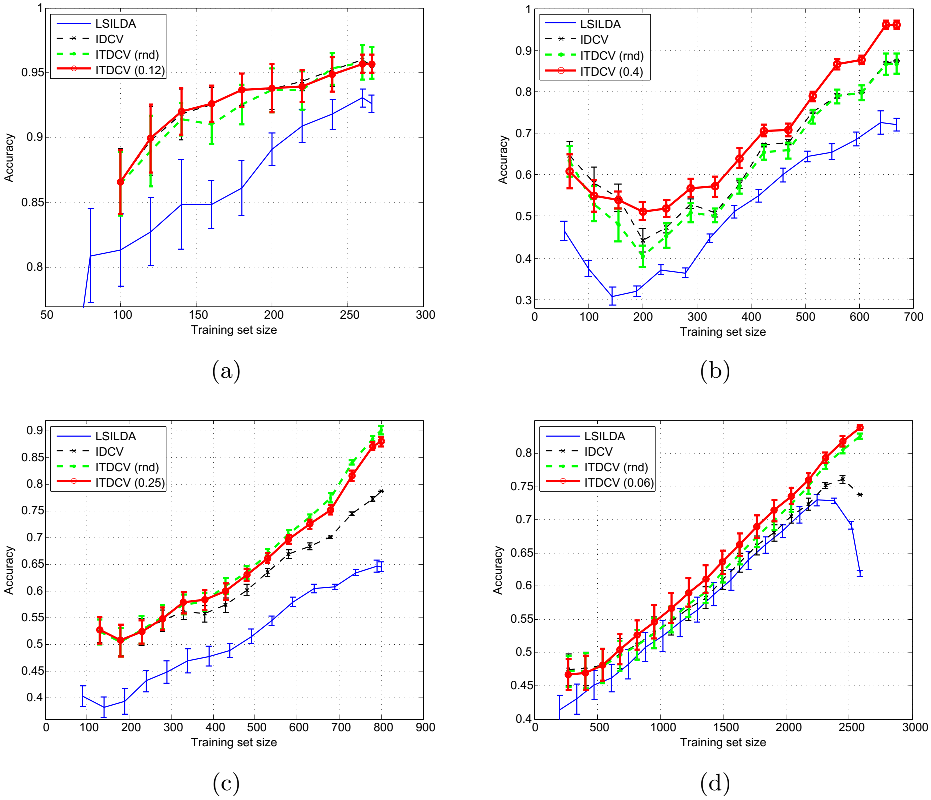
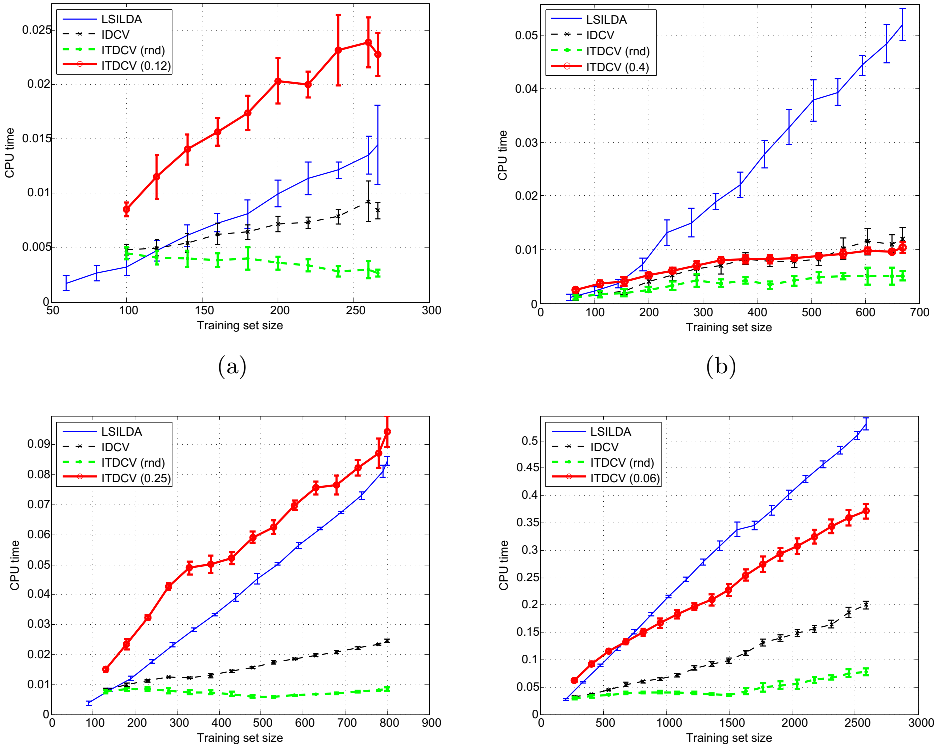


Fig. 1. Recognition rates obtained using the proposed method (ITDCV), the same with random selection, the LS-ILDA and the IDCV algorithm on the 4 of the databases considered: (a) ORL, (b) Yale, (c) AR, and (d) CMU-PIE

method will be referred to as Incremental Truncated DCV (ITDCV). The LS-ILDA algorithm [12] has been considered as baseline. Moreover, the QRD-based IDCV that randomly discards directions has been considered and referred to as ITDCV(rnd). The ν parameter has been tuned manually for each dataset, and the ITDCV(rnd) has been adjusted in such a way that the number of directions discarded is as close as possible to the one with the tuned ITDCV(ν).

In this work, 4 publicly available image databases have been considered as a convenient and widely used case of high dimensional data which leads to an undersampled situation. Images have been normalized in intensity and roughly aligned. These databases are ORL (d=1024, M=400, c=40), MNIST (d=1024, M=1000, c=10), COIL (d=1600, M=1200, c=40) and CMU-PIE (d=2700, M= 3808, c=68). Further details are given in previous similar studies [14]. An experimental setup in which more training data becomes available to the algorithm has been designed. In particular and for each database, the available data has



(a)

(b)

(c)

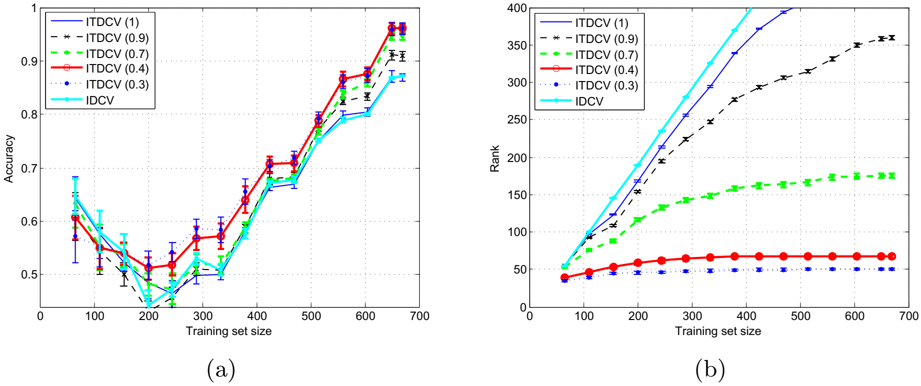
(d)

Fig. 2. CPU times (in seconds) obtained using the proposed method (ITDCV), the same with random selection, the LS-ILDA and the IDCV algorithm on the 4 of the databases considered: (a) ORL, (b) MNIST, (c) COIL, and (d) CMU-PIE

been split into 2 disjoint sets. The first 2/3 of the data is used for training and the rest is kept for test. This is repeated 10 times. The results presented correspond then to an average across the 10 runs along with corresponding standard deviations. The accuracy of the Nearest Neighbor classifier (with $k = 1$) in the projected subspace has been considered as a performance measure [5, 12]. Also, CPU times for each algorithm at each iteration have been measured.

Figure 1 shows that the proposed ITDCV gives in all cases the best or as good as the best performance results. It is worth noting that IDCV gives also the best results for ORL and MNIST but are significantly worse for COIL and CMU-PIE. Another more surprising fact is that randomly discarding directions in the same amount lead to equally good results in ORL and COIL. LS-ILDA gives always worst results than IDCV as it was already shown [14]. In the largest database, CMU-PIE, it can be observed that the performance of both IDCV and LS-ILDA deteriorates as the value of M approaches d.

On the other hand, CPU times corresponding to incrementally updating projections shown in Figure 2 illustrate the fact that the proposed algorithm behaves linearly with M. Observe that the time spent is worst than the one from IDCV but moderately good as compared to the one from LS-ILDA. ITDCV(rnd) is obviously the best in terms of computational burden as it does only a fraction of the job done by IDCV. To better understand how ITDCV behaves, Figure 3 shows both accuracies and rank values, r, corresponding to ITDCV(ν) for several values of ν in the case of MNIST. The IDCV algorithm is also shown for comparison purposes. It can be seen that the value of ν in the proposed algorithm has a very moderate impact on its performance that is always as good or better than IDCV. On the other hand, the parameter has a strong impact in rank growth and, consequently, in computational time. In fact, we see that it is possible to significantly improve the computing times in Figure 2 without significantly degrading the performances in Figure 1.



(a) (b)

Fig. 3. Proposed ITDCV method for different values of the parameter ν compared to the exact IDCV method: (a) accuracy, (b) within-class scatter rank

5 Concluding Remarks

An approximate incremental algorithm to compute DCVs and corresponding subspaces has been proposed. The algorithm is based in a rank one SVD update along with a restriction on the growth of the range space of the within-class scatter matrix. The algorithm is very efficient and numerically stable and gives the same or better performance results than the DCV algorithm. Very competitive results both in performance and complexity when compared to one of the best incremental LDA implementations to date has been obtained in the empirical evaluation carried out. Further work is being done on applying this algorithm along with an appropriate tunning strategy for more realistic problems of significantly larger sizes.

References

1. Fukunaga, K.: Introduction to statistical pattern recognition, 2nd edn. Academic Press Professional, Inc., San Diego (1990)
2. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
3. Wang, X., Tang, X.: A unified framework for subspace face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 26(9), 1222–1228 (2004)
4. Yu, H., Yang, J.: A direct lda algorithm for high-dimensional data – with application to face recognition. Pattern Recognition 34(10), 2067–2070 (2001)
5. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 27(1), 4–13 (2005)
6. Ye, J.: Least squares linear discriminant analysis. In: ICML 2007: Proc. of the 24th Intl. Conf. on Machine Learning, pp. 1087–1093. ACM, New York (2007)
7. Chandrasekaran, S., Manjunath, B., Wang, Y., Winkler, J., Zhang, H.: An eigenspace update algorithm for image analysis. Graphical Models and Image Processing 59(5), 321–332 (1997)
8. Ozawa, S., Toh, S.L., Abe, S., Pang, S., Kasabov, N.: Incremental learning of feature space and classifier for face recognition. Neur. Netw. 18(5), 575–584 (2005)
9. Ye, J., Li, Q., Xiong, H., Park, H., Janardan, R., Kumar, V.: Idr/qr: An incremental dimension reduction algorithm via qr decomposition. IEEE Trans. on Knowl. and Data Eng. 17(9), 1208–1222 (2005)
10. Kim, T.K., Wong, S.F., Stenger, B., Kittler, J., Cipolla, R.: Incremental linear discriminant analysis using sufficient spanning set approximations. In: Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8 (2007)
11. Zhao, H., Yuen, P.C.: Incremental linear discriminant analysis for face recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 38(1), 210–221 (2008)
12. Liu, L.P., Jiang, Y., Zhou, Z.H.: Least square incremental linear discriminant analysis. In: Intl Conf. on Data Mining, ICDM 2009, pp. 298–306 (2009)
13. Chu, D., Thye, G.S.: A new and fast implementation for null space based linear discriminant analysis. Pattern Recognition 43(4), 1373–1379 (2010)
14. Ferri, F.J., Diaz-Chito, K., Díaz-Villanueva, W.: Efficient dimensionality reduction on undersampled problems through incremental discriminative common vectors. In: Intl. Conf. on Data Mining Workshops, ICDMW 2010, pp. 1159–1166 (2010)
15. Lu, G.F., Zou, J., Wang, Y.: Incremental learning of discriminant common vectors for feature extraction. Appl. Math. and Computation 218(22), 11269–11278 (2012)
16. Golub, G.H., Van Loan, C.F.: Matrix Computations (Johns Hopkins Studies in Mathematical Sciences), 3rd edn. The Johns Hopkins Univ. Press (1996)
17. Lu, G.F., Zheng, W.: Complexity-reduced implementations of complete and null-space-based linear discriminant analysis. Neural Networks (to appear, 2013)
18. Tamura, A., Zhao, Q.: Rough common vector: A new approach to face recognition. In: IEEE Intl. Conf. on Syst, Man and Cybernetics, pp. 2366–2371 (2007)
19. Brand, M.: Incremental singular value decomposition of uncertain data with missing values. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 707–720. Springer, Heidelberg (2002)

Centering SVDD for Unsupervised Feature Representation in Object Classification

Dong Wang and Xiaoyang Tan*

Department of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics,
#29 Yudao Street, Nanjing 210016, P.R. China
x.tan@nuaa.edu.cn

Abstract. Learning good feature representation from unlabeled data has attracted researchers great attention recently. Among others, K-means clustering algorithm is popularly used to map the input data into a feature representation, by finding the nearest centroid for each input point. However, this ignores the density information of each cluster completely and the resulting representation may be too terse. In this paper, we proposed a SVDD (Support Vector Data Description) based method to address these issues. The key idea of our method is to use SVDD to measure the density of each cluster resulted from K-Means clustering, based on which a robust feature representation can be derived. For this purpose, we add a new constraint to the original SVDD objective function to make the model align better with the data. In addition, we show that our modified SVDD can be solved very efficiently as a linear programming problem, instead of as a quadratic one. The effectiveness and feasibility of the proposed method is verified on two object classification databases with promising results.

Keywords: Feature learning, K-means, Support Vector Data Description(SVDD), C-SVDD, object classification.

1 Introduction

Learning good feature representation from unlabeled data is the key to make progress in recognition and classification tasks, and has attracted great attention and interest from both academia and industry recently [1]. Deep learning method which aims to learn multiple layers of abstract representations from data has gained much success and has become a popular way for representation learning. In this method layers of representation is usually obtained by greedily training one layer at a time on the lower level [2], [3], [4], using an unsupervised learning algorithm. In this sense, the performance of single-layer learning has an big effect on the final representation. Neural network based single-layer methods, such as

* This work was supported by the National Science Foundation of China (61073112, 61035003,61373060), Jiangsu Science Foundation (BK2012793), Qing Lan Project, and Research Fund for the Doctoral Program (RFDP) (20123218110033).

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 376–383, 2013.

autoencoder [5] and RBM (Restricted Boltzmann Manchine,[6]), are widely used for this but they have the disadvantages that the models are usually very complex and have many parameters to adjust. In addition, many parameters involved are need to be set through cross-validation, which is very time-consuming.

That is why a simple and fast method is preferred for unsupervised feature learning. Among others K-means clustering algorithm is commonly used to map the input data into a feature representation. The simplest way for this is to map each data point to its nearest cluster center and use it as the feature to describe the data. There is only one parameter involved in the K-means based method, i.e., the number of clusters, hence the model is very simple and fast. Coates et al. [7] shows that the K-means based encoder achieves the best performance compared with sparse autoencoder, sparse RBM and GMM (Guassian Mixture Model) under some circumstances. Despite of the success, the above K-means based feature representation scheme is not perfect from the aspect of the richness of information it conveys. Actually, such a representation is too terse, and does not take the non-uniform distribution of cluster size into account. Intuitively, those clusters containing more data are likely to be part of the features with higher influential power, compared to the smaller ones.

In this paper, we proposed a SVDD (Support Vector Data Description, [8], [9]) based method to address these issues. The key idea of our method is to use SVDD to measure the density of each cluster resulted from K-means clustering, based on which more robust feature representation could be built. Actually the K-means algorithm lacks a robust definition of the size of its clusters, since the nearest center principle is not robust against the noise or outliers common in real world applications. We advocate that the SVDD could be a good way to address this issue. Actually SVDD is a widely used tool to find a minimal a closed spherical boundary to include all the data belong to target class and therefore, given a cluster of data, we are expecting SVDD to generate a ball containing the all normal data excepting outliers. Performing this procedure on all the clusters of K-means, we will finally get K SVDD balls on which our representation can be built. In addition, considering that a bigger ball is more influential than smaller ones, we use the distance from the data to each ball's surface instead of the center as the feature.

One problem of our model comes from the instability of SVDD's center, due to the fact that its position is mainly determined by the support vectors on the boundary and the noise in the data may deviate the center far from the mode (c.f., Fig. 3(left)). This makes the SVDD ball not be consistent with the data's distribution when used for feature representation. To address this, we add a new constraint to the original SVDD objective function to make the model align better with the data. In addition, we show that our modified SVDD can be solved very efficiently as a linear programming problem, instead of as a quadratic one. Experiments on the AR face dataset and CIFAR-10 object database show that it is robust, efficient, and when combined with K-means, it provides a much richer representation for the input data and thus improves the performance of object classification.

2 Preliminaries

2.1 Unsupervised Feature Learning

The overall pipeline of the feature representation is as follows. For a given image, a set of patches are first sampled at the positions of a regular grid [7]. By mapping those patches to their nearest cluster centers, a set of feature maps could be obtained. Then one can pooling on these and reshape them into a vector, which yields the final feature representation for the input image. It is worthy mentioning that there is a small difference between the above method and others such as the CNN network [10], [11], i.e., instead of using a learnt filtering bank for convolution, the K-means centers are used as references for feature mapping. In other words, the cluster centers play the same role as the filtering bank in CNN network but its way for feature mapping is different from the latter.

2.2 K-Means for Feature Learning

K-means is a data clustering algorithm to divide data into a set of K clusters, with Euclidean distance as similarity measure. It aims to minimize the sum of distance between all data to their corresponding centers. Let X=$\{x_i\}$,i=1,...,n be the set of n d-dimensional points, $C = \{c_k\}$, k=1,...,K be the K clusters. Let μ_k be the mean of the cluster c_k. The objective function is defined as: $J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$.

As mentioned in the previous section, each cluster would be used to produce a feature mapping. So if we have K clusters, the dimension of the resulting feature representation will be K as well. The simplest way for feature mapping is the so-called "hard coding" method, i.e., simply setting the winner cluster center on while all the others off, as follows,

$$f_k(x) = \begin{cases} 1 & \text{if } k = argmin_j \|c_j - x\|_2^2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The resulting K-dimensional vector f can be thought as the MAP estimate of the input point x given the K-means model. However it is too sparse and is often not representative of the full posterior mass. A better summary is the following "soft coding":

$$f_k(x) = max\{0, \mu(z) - z_k\} \tag{2}$$

where $z_k = \|x - c_k\|_2$, and $\mu(z)$ is the mean of the elements of z. This activation function outputs 0 for the feature f_k that have an above average distance to the centroid c_k. This model leads to a less sparse representation (roughly half of the features are found to be 0 in our experiments), but as shown in the experimental section, it significantly improves the classification performance.

However, this method does not take the characteristics of each cluster into consideration. Actually, the number of data point in each cluster is usually different, so is the distribution of data points in each cluster. We believe that these differences would make a difference in feature representation as well. However,

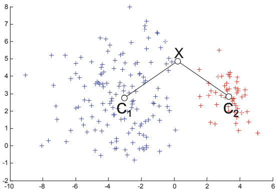


Fig. 1. Illustration of the unequal cluster effect

the aforementioned K-means feature mapping scheme completely ignores these and only use the position of center for coding. As shown in Fig. 1, although the data point x has the same distance to the centers $C1$ and $C2$ of two clusters, it should be assigned a higher score on $C1$ than on $C2$ since the former cluster $C1$ is much bigger than the latter. In practice such unequal clusters are not uncommon and the K-means method by itself can not reliably grasp the size of its clusters due to the existence of outliers. To this end, we propose an SVDD based method to describe the density and distribution of each cluster and use this for more robust feature representation.

3 The Proposed Method

3.1 Using SVDD Ball to Cover Unequal Clusters

Assume that a data set contains N data objects, $\{x_i\}$, $i = 1, ..., n$ and a ball is described by its center a and the radius R. The goal of SVDD (Support Vector Data Description, [8]) is to find a closed spherical boundary around the given data points. In order to avoid the influence of outliers, SVDD actually faces the tradeoff between two conflicting goals, i.e., minimizing the radius while covering as many data points as possible.

The SVDD method can be understood as a type of one-class SVM and its boundary is solely determined by support vectors points. SVDD allows us to summarize a group of data points in a nice and robust way. Hence it is natural to use SVDD ball to model each cluster from K-means, thereby combining the strength of both models. In particular, for a given data point we first compute its distance h_k to the surface of each SVDD ball C_k, and then use the following soft coding method for feature representation similar to E.q.(2): $f_k(x) = max\{0, g(z) - h_k\}$, where $g(z) = \mu(z) - \mu(R)$ and $\mu(R)$ is the mean of radius R of balls, while $h_k = |z_k - R_k|$ is the distance from the point to the surface of the SVDD ball.

Shown in Fig. 2 for a data point x, C_i, i=1,2 respectively are the centroids of two SVDD balls with R_i, i=1,2 being their the radius respectively, and $h_i = |z_i - R_i|$ is the distance from x to the surface of $i - th$ ball. Since the distances from x to $C1$ and C_2 are equal, x will get the same scores on the two ball with the K-means scheme (c.f., E.q.(2)). However, if we take the density and size of the clusters into accounts, the score from $C2$ should be higher and that is exactly our method does.

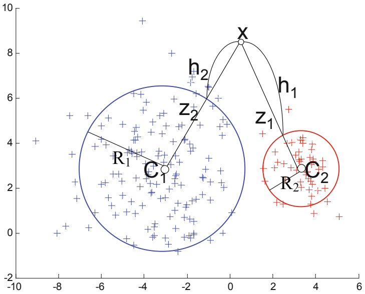


Fig. 2. Using the SVDD ball to cover the clusters of K-means

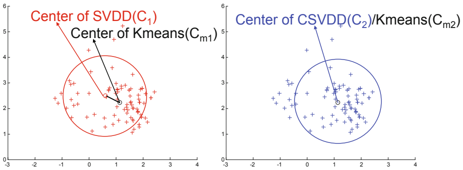


Fig. 3. Illustration of the difference between SVDD and C-SVDD, where the left ball ($C1$) is from SVDD while the right (C2) is from C-SVDD. Note that the center of C-SVDD ball aligns better with the high density region of the data points. The C_m marks the center of K-means.

3.2 The C-SVDD Model

Although SVDD ball provides a robust way to describe the cluster of data, one unwelcome property of the ball is that it may not align well with the distribution of data points in that cluster. As illustrated in Fig. 3 (left), although the SVDD ball covers the cluster $C1$ well, its center is biased to the region with low density. This should be avoided since it actually gives suboptimal estimates on the distribution of the cluster of data.

To address this issue, inspired by the observation that the centers of K-means are always located at the corresponding mode of their local density, we propose to shift the SVDD ball to the centroid of the data such that it may fit better with the distribution of the data in a cluster. Our new objective function is then formulated as follows,

$$\begin{aligned} &min_{R,\xi_i}\ R^2 + C\sum_{i=1}^{N}\xi_i \\ &s.t.\ \|x_i - a\|^2 \leq R^2 + \xi_i \\ &a = \frac{1}{N}\sum_{i=1}^{N} x_i \\ &\xi_i \geq 0 \end{aligned} \tag{3}$$

where $\|.\|$ is the L_2-norm and ξ_i is the slack variable to the ith sample x_i. With Lagrange multipliers $\alpha_i \geq 0$ and $\alpha_j \geq 0$ according to KKT Conditions, one has the following dual function:

$$\begin{aligned} & max \sum_i \alpha_i \langle x_i, x_i \rangle - \frac{2}{N} \sum_i \sum_j \alpha_i \langle x_i, x_j \rangle \\ & s.t. \sum_i \alpha_i = 1 \ , \ \alpha_i \in [0, C] \ , \ i = 1, ..., N \end{aligned} \tag{4}$$

Eq.(4) can be rewritten as:

$$\begin{aligned} & min \ \frac{2}{N} \alpha^T He - \alpha^T F \\ & s.t. \ \alpha^T e = 1 \ , \ \alpha_i \in [0, C] \ , \ i = 1, ..., N \end{aligned} \tag{5}$$

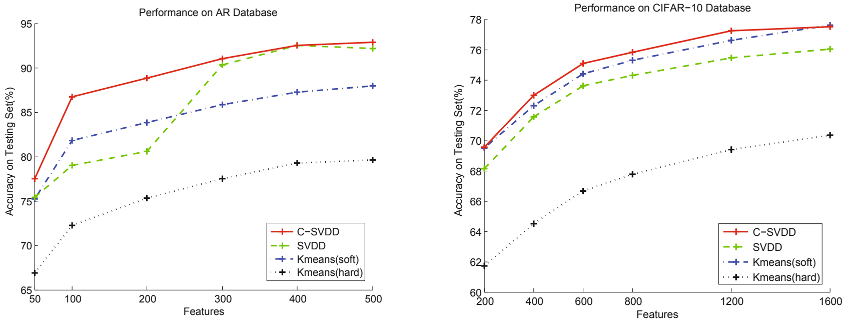
where $H = (\langle x_i, x_j \rangle))_{N \times N}$, $F = (\langle x_i, x_i \rangle)_{N \times 1}$, $e = (1, 1, ..., 1)^T$. It is worthy mentioning that this objective function is linear to α, and thus can be solve efficiently with a linear programming algorithm.

Since the model is centered towards the mode of the distribution of the data points in a cluster, we named our method as C-SVDD (centered-SVDD). Figure.3 shows the difference between SVDD and C-SVDD, where the left result is from SVDD and the right from C-SVDD. We can see that our new model aligns better with the density of the data points, as expected.

4 Experiments and Analysis

To investigate whether the proposed method can produce good feature representation. We conducted a series of experiments on the AR face database [12] and the CIFAR-10 object dataset [13], on each of which, we compared our method (C-SVDD with K-means) with other three types of feature mapping strategies, i.e., K-means(hard), K-means(soft) and SVDD (combined with K-means). All the images in use undergone whitening preprocessing before being sampled for feature mapping [7].

The AR face database [12] contains over 4,000 color images corresponding to 126 people's faces. Every person has 2 sessions images with 13 for each. Images are all frontal view faces with different facial expressions, illumination conditions, and occlusions. Here we use all the images from the first session for training while those in the second session for testing. All images are resized to 64×64. For training we sample 40000 patches with size 6×6 from training set, and cluster them using K-means by varying the number of clusters K. Then we do the feature mapping as described in the previous section. Note that for the normalization parameter C in SVDD and C-SVDD, If $C = 0$, the representation result of $C - SVDD$ is equal to K-means, while a larger C value means more noise is allowed to enter the ball. We use 5-cross validation to set its value from a range of {0.005, 0.01, 0.1, 1}.



(a) (b)

Fig. 4. Comparative performance of the proposed method with various K-means based encoding strategies on (a) the AR dataset and (b) the CITAR-10 dataset

The CIFAR-10 [13] dataset is a more complicated database which consists of 60000 32x32 color images in 10 classes with 6000 images per class. There are 50000 training images and 10000 test images, and the training set is divided into five batches. We also use 5-cross validation to set the best C value for C-SVDD and SVDD with a range of $\{0.004, 0.005, 0.006, 0.008, 0.01\}$. The receptive field is 6 by 6, and 400000 patches are sampled for training.

Figure 4 gives the results. It can be seen that our C-SVDD-based representation method is the best performer on both datasets. The K-means (hard) method is the worst one as expected due to its extremely sparse representation, while replacing the hard coding with a soft one (K-means (soft)) significantly improves the performance. The figure also reveals that the scheme of simply adding SVDD ball onto the top of soft K-means does not necessarily work and may actually hurt the performance due to the bias it introduced (as explained in the previous section). However, once this problem solved, the performance is improved a lot. Another point needing to be pointed out is that when the number of features (i.e., the cluster number K in K-means) increases, the performance of all the four methods improves consistently. This indicates the importance of encoding richer information in the feature representation.

Table 1 gives the comparative performance (%) of our method with other state-of-the-art single-layer network results on the CIFAR-10 dataset. For a fair comparison, we adopted the same evaluation protocol as that in [7], and all the results except the last row are directly cited from it. It is clear that our C-SVDD method performs the best among the compared methods.

5 Conclusion

In this paper, we proposed a SVDD based feature learning algorithm that enhances the K-means ”soft” feature representation. The key idea of our method is to describe the density and distribution of each cluster from K-means with a

Table 1. Comparative performance (%) with other state-of-the-art single-layer network methods on the CIFAR-10 dataset

Algorithm	Accuracy
Sparse auto-encoder [7]	73.4
Sparse RBM [7]	72.4
K-means (Hard) [7]	68.6
K-means (Triangle, 4000 features) [7]	79.6
C-SVDD (4000 features) (ours)	**79.8**

SVDD ball for more robust feature representation. For this purpose, we presented a new SVDD algorithm called C-SVDD that centers the SVDD ball towards the mode of local density of each cluster. Furthermore we show that the objective of C-SVDD can be solved very efficiently as a linear programming problem. Experiments on the AR and the CIFAR-10 database show that our C-SVDD based feature representation method outperforms the original K-means based scheme.

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. arXiv preprint arXiv:12065538 (2012)
2. Le, Q.V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G.S., et al.: Building high-level features using large scale unsupervised learning. arXiv preprint arXiv:11126209 (2011)
3. Agarwal, A., Triggs, B.: Hyperfeatures – multilevel local coding for visual recognition. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 30–43. Springer, Heidelberg (2006)
4. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153. IEEE (2009)
5. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)
6. Cueto, M.A., Morton, J., Sturmfels, B.: Geometry of the restricted Boltzmann machine. In: Viana, M., Wynn, H. (eds.) Algebraic Methods in Statistics and Probability. AMS, Contemporary Mathematics, vol. 516, pp. 135–153 (2010)
7. Coates, A., Lee, H., Ng, A.Y.: An analysis of single-layer networks in unsupervised feature learning. Ann Arbor 1001, 48109 (2010)
8. Tax, D.M., Duin, R.P.: Support vector data description. Machine Learning 54(1), 45–66 (2004)
9. Xu, J., Yao, J., Ni, L.: Fault detection based on SVDD and cluster algorithm. In: 2011 International Conference on Electronics, Communications and Control (ICECC), pp. 2050–2052. IEEE (2011)
10. Niu, X.X., Suen, C.Y.: A novel hybrid CNN–SVM classifier for recognizing handwritten digits. Pattern Recognition 45(4), 1318–1325 (2012)
11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
12. Martinez, A.M.: The AR face database. CVC Technical Report 24 (1998)
13. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009)

Improved HOG Descriptors in Image Classification with CP Decomposition

Tan Vo[1], Dat Tran[1], Wanli Ma[1,2], and Khoa Nguyen[1]

[1] Faculty of Education, Science, Technology & Mathematics
University of Canberra, ACT 2601, Australia
[2] Department of Computer Science, University of Houston Downtown, USA
{tan.vo,dat.tran,wanli.ma,khoa.nguyen}@canberra.edu.au

Abstract. Histogram of Oriented Gradients (HOG) has been widely used in computer vision as feature descriptors for detecting objects in scenes. We present in this paper a new approach to HOG in image classification that will provide an opportunity to explore new ways to improve the effectiveness of HOG image descriptors. We investigate applying tensor decomposition on HOG descriptors then using them as image features to build image models using support vector machine. The aim of this approach is to produce a more robust and compact version of HOG features. An image classification experiment is performed to evaluate the effectiveness of this approach as well as to identify all ideal parameter values involved. Experimental results show a good improvement in image classification rate for the proposed approach.

Keywords: HOG, tensor, CP decomposition, Image Classification, Support Vector Machine.

1 Introduction

In modern computer vision, the mechanism that represents the visual features of an image entity is known as image descriptor. The most commonly used image descriptors are HOG [1] and Scale-invariant feature transform (SIFT) [2]. These descriptors share the same concept of implementation, in the sense that the local histograms of orientation of patches across the image are used to produce the vector. The main advantage of these descriptors are their invariance nature to deformations of object within images: rotation, illumination, scale, viewpoint, noise, etc [2]. Being quite similar in the implementation, the factor that separates these descriptors sometimes boil down to the way they are being used.

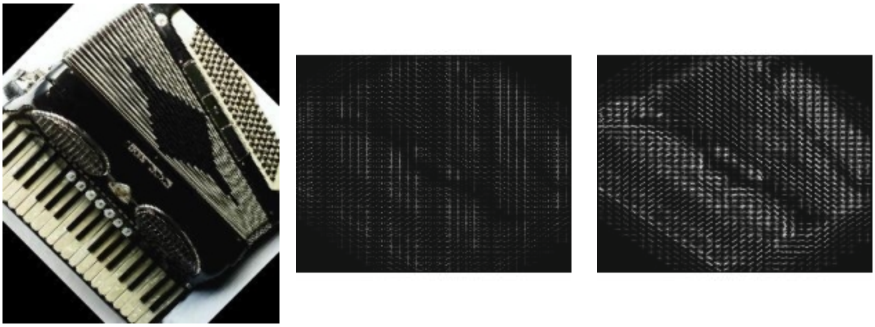
In Dalal and Triggs's original work [1], HOG descriptors are used as feature vectors for a linear support vector machine (SVM), which performs like a sliding window human detector within scenes. Since the work of Felzenszwalb et al. [3], HOG is now also known in providing robust and effective features to be used for object detection.

In many cases, HOG descriptors can be processed and utilized in similar fashion to SIFT descriptors. In particular, this paper considers using HOG as a

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 384–391, 2013.

feature extracting mechanism for the image classification task. Likewise to SIFT, HOG descriptors allow an image classifier to produce matches on a given image to the images that were used to train the image classifier by constructing a visual bag-of-words (BoW) model of the HOG descriptor and using it in tandem with a machine learning algorithm such as SVM.

Due to the nature they are created, HOG descriptors are generally represented in three-dimensional space and, depending on the resolution of the image, also tends to be high in volume. Figure 1 visualizes the HOG descriptors created on the image of an accordion. Unlike SIFT, a value of O, number of orientations, can be configured in HOG operation and this value can influent the features quite significantly (as shown in Figure 1b and Figure 1c). This flexibility allows HOG to be used for a variety of image detection problems.



(a) Original image (b) $O = 3$ orientations (c) $O = 21$ orientations

Fig. 1. Visual representation of HOG descriptors

In recent years, there exists a trend in which researchers have been using techniques such as PCA (principal component analysis) and SVD (singular value decomposition) on these types of descriptors. The goal is to either lessen the problem complexity by dimensionality reduction of the original descriptors [4] or to produce improved version of themselves [5], which generally results in more robustness in recognition and classification. This trend has shown potential in the work of compacting and reducing these image descriptors, which motivates us to propose a new approach, in which canonical polyadic (CP) decomposition [6,7] is applied to HOG descriptor before they are being used for image classification. CP decomposition is essentially a generalization of matrix SVD to tensors, or multidimensional array.

2 Proposed Method

2.1 HOG

Figure 2 summaries the steps in extracting the HOG descriptors of an image. This process begins with dividing a given image into cells of equal size as in 2(a).

Within each cell, a histogram of gradient directions, or edge orientations, is accumulated over the pixels as in 2(b). The orientation $\theta(x, y)$ and the magnitude $r(x, y)$ of a pixel (x, y) are calculated with a 1-D discrete derivations mask $[-1, 0, 1]$ and its transpose $[-1, 0, 1]^\top$. The magnitude $r(x, y)$ is calculated with the color channel with the largest gradient magnitude. Let $O = 9$ be the number of orientations, there will be $2 \times 9 = 18$ directed orientation bins allocated, or one bin for every $20°$ in the range $0° - 360°$: 2 orientations ($\pm$) for each of the 9 undirected gradient directions (Dalal et al. [1]).

The next step in HOG is block normalization, in which blocks are generated by grouping four adjacent cells together (sliding of each cell) as visualized in 2(c). Let vector v be the stacking of the positive direction histogram in a block, $\|v\|_2$ be the two-norm of v and ϵ a very small number (it is presumed that ϵ also has an insignificant value), the norm (l^2-norm) of a block is defined as

$$v = v \Big/ \sqrt{\|v\|_2^2 + \epsilon^2} \tag{1}$$

The final step 2(d) produces the actual descriptors. For each cell, four normalization factors can be obtained as the inverse of the norm of the four blocks that contain it. Four copies of the cell's undirected 9-dimensioned histogram will then be normalized with each normalization factor, separately. The results are stacked and clipped at 0.2. The process will produce a vector of $4 \times 9 = 36$ in length. This is used as the HOG descriptor representing the cell.

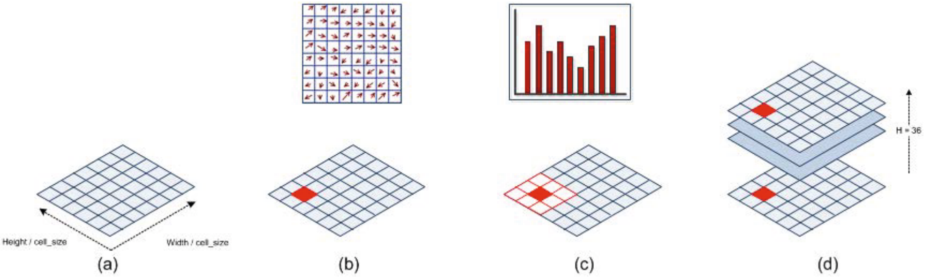


Fig. 2. Process of creating HOG descriptors

Another HOG variant is considered in this paper is UoC/TTI [3]. This type of descriptors are created with a slightly different process, in which *a*) the normalization is performed over both directed (18 bins) and undirected histograms (9 bins), i,e produce a vector of $4 \times (2 + 1) \times 9$ in length ; *b*) the dimensionality of the result is reduced with a PCA variation to the length of $(2 + 1) \times 9$; and *c*) the l^1norm of the four normalized undirected histograms is computed and stored as additional four dimensions . With that, the UoC/TTI process produces 31-dimensioned descriptors ($(2 + 1) \times 9 + 4$), rather than 36 in the case of Dalal-Triggs variant with the UoC/TTI variant.

In Figure 2(d), as blocks are visited from left to right and top to bottom, they form the final descriptor of the image. The HOG descriptors are normally structured in the form of a 3-dimensional array, or third-order tensor. Let cH and cW be the numbers of cells along the image's height and width, respectively. For the Dalal-Triggs variant, the HOG tensor has a dimension of $cH \times cW \times 36$ whereas it has the dimension of $cH \times cW \times 31$.

2.2 CP Tensor Decomposition

The process of CP decomposing a tensor involves factorizing it into a sum of component rank-one tensors. In this situation, given a third-order tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and a positive integer tensor-rank R, this process is denoted as:

$$\boldsymbol{X} \approx [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}]\!] = \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \mathbf{a}_r^{(3)} \tag{2}$$

where operator $\circ$ is the vector outer product, with $\mathbf{a}_r^{(n)} \in \mathbb{R}^{I_n}$. Factor matrices $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$ and $\mathbf{A}^{(3)}$ are combinations of those rank-one components $\mathbf{A}^{(n)} = \left[\mathbf{a}_1^{(n)} \, \mathbf{a}_2^{(n)} \cdots \mathbf{a}_R^{(n)} \right]$. Core vector $\boldsymbol{\lambda} \in \mathbb{R}^R$ is used to normalize the columns of factor matrices to length one. Equation 2 can be re-written in another form with these statements:

$$\begin{aligned}
\mathbf{X}^{(1)} &\approx \mathbf{A}^{(1)} \operatorname{diag}(\boldsymbol{\lambda})(\mathbf{A}^{(3)} \odot \mathbf{A}^{(2)})^\top, \\
\mathbf{X}^{(2)} &\approx \mathbf{A}^{(2)} \operatorname{diag}(\boldsymbol{\lambda})(\mathbf{A}^{(3)} \odot \mathbf{A}^{(1)})^\top, \\
\mathbf{X}^{(3)} &\approx \mathbf{A}^{(3)} \operatorname{diag}(\boldsymbol{\lambda})(\mathbf{A}^{(2)} \odot \mathbf{A}^{(1)})^\top,
\end{aligned} \tag{3}$$

where the operator $\odot$ denotes a Khatri-Rao product. This column-wise Kronecker product of two matrices $\mathbf{A} = [\,\mathbf{A}_1 \mid \mathbf{A}_2 \mid \ldots \mid \mathbf{A}_n\,]$ and $\mathbf{B} = [\,\mathbf{B}_1 \mid \mathbf{B}_2 \mid \ldots \mid \mathbf{B}_n\,]$, where $\mathbf{A}^n$ and $\mathbf{B}^n$ are column vectors of $\mathbf{A}$ and $\mathbf{B}$, is defined as follows:

$$\boldsymbol{A} \odot \boldsymbol{B} = \left[\begin{array}{c|c|c|c}
\boldsymbol{A}_1^1 \boldsymbol{B}^1 & \boldsymbol{A}_1^2 \boldsymbol{B}^2 & \ldots\ldots & \boldsymbol{A}_1^n \boldsymbol{B}^n \\
\vdots & \vdots & \ddots & \vdots \\
\boldsymbol{A}_m^1 \boldsymbol{B}^1 & \boldsymbol{A}_m^2 \boldsymbol{B}^2 & \ldots\ldots & \boldsymbol{A}_m^n \boldsymbol{B}^n
\end{array} \right] \tag{4}$$

The decomposition process of tensor $\boldsymbol{X}$ is to identify a composition $\boldsymbol{X}'$ such that it satisfy the condition:

$$\min_{\boldsymbol{X}'} \left\| \boldsymbol{X} - \boldsymbol{X}' \right\| = \min_{\mathbf{A}^{(i)}} \left\| \boldsymbol{X} - [\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}] \right\| \tag{5}$$

or in component form:

$$\begin{aligned}
\min_{\hat{\mathbf{A}}^{(1)}} &= \left\| \boldsymbol{X}^{(1)} - \hat{\mathbf{A}}^{(1)}(\mathbf{A}^{(3)} \odot \mathbf{A}^{(2)})^\top \right\|, \\
\min_{\hat{\mathbf{A}}^{(2)}} &= \left\| \boldsymbol{X}^{(2)} - \hat{\mathbf{A}}^{(2)}(\mathbf{A}^{(1)} \odot \mathbf{A}^{(3)})^\top \right\|, \\
\min_{\hat{\mathbf{A}}^{(3)}} &= \left\| \boldsymbol{X}^{(3)} - \hat{\mathbf{A}}^{(3)}(\mathbf{A}^{(2)} \odot \mathbf{A}^{(1)})^\top \right\|
\end{aligned} \tag{6}$$

where the column vector normalization of $\mathbf{A}^{(n)}$ is defined as $\hat{\mathbf{A}}^{(n)} = \mathbf{A}^{(n)}.\operatorname{diag}(\boldsymbol{\lambda})$. Given $\boldsymbol{A}^\dagger$ that represents the pseudoinverse of a matrix $\boldsymbol{A}$ [8], the pseudoinverse of a Khatri-Rao product has this property [8]

$$(\boldsymbol{A} \odot \boldsymbol{B})^\dagger = (\boldsymbol{A}^\top \boldsymbol{A} * \boldsymbol{B}^\top \boldsymbol{B})^\dagger (\boldsymbol{A} \odot \boldsymbol{B})^\top, \tag{7}$$

where the matrix operator "*" represents the Hadamard element-wise matrix product [8]. Equipped with Property 7, an optimal solution $\hat{\mathbf{A}}^{(1)}$ for Formula 6 can then be written as

$$\begin{aligned}
\hat{\mathbf{A}}^{(1)} &= \boldsymbol{X}^{(1)} \left[(\mathbf{A}^{(3)} \odot \mathbf{A}^{(2)})^\top \right]^\dagger \\
&= \boldsymbol{X}^{(1)}(\mathbf{A}^{(3)} \odot \mathbf{A}^{(2)})(\mathbf{A}^{(3)^\top}\mathbf{A}^{(3)} * \mathbf{A}^{(2)\top}\mathbf{A}^{(2)})^\dagger,
\end{aligned} \tag{8}$$

Similar calculations then can be applied to calculate $\hat{\mathbf{A}}^{(2)}$ and $\hat{\mathbf{A}}^{(3)}$ in Equation 8. In this paper, the core of the CP decomposition process is the alternating least squares (ALS) algorithm [6,7]. Figure 3 describes the steps involved in ALS: Using Equation 8, it fixes $\mathbf{A}^{(\mathbf{2})}$ and $\mathbf{A}^{(\mathbf{3})}$ to solve for $\mathbf{A}^{(\mathbf{1})}$, then fixes $\mathbf{A}^{(\mathbf{3})}$ and $\mathbf{A}^{(\mathbf{1})}$ to solve for $\mathbf{A}^{(\mathbf{2})}$, then fixes $\mathbf{A}^{(\mathbf{1})}$ and $\mathbf{A}^{(\mathbf{2})}$ to solve for $\mathbf{A}^{(\mathbf{3})}$, and continues to repeat the entire procedure until some convergence criterion is satisfied. The result of the ALS processes with a R-ranked on a third-order tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ includes a core tensor $\boldsymbol{\lambda}$ as well as three tensors: $\boldsymbol{A}_{(1)} \in \mathbb{R}^{I_1 \times R}$, $\boldsymbol{A}_{(2)} \in \mathbb{R}^{I_2 \times R}$ and $\boldsymbol{A}_{(3)} \in \mathbb{R}^{I_3 \times R}$.

2.3 Feature of Image

The proposed approach involves applying CP decomposition with an R-ranked onto the HOG descriptors. Depending on the HOG variants as well as the number of orientations, we construct D-dimensional descriptors. The HOG descriptors of an image, in the form of $cH \times cW \times d$ tensor that will be decomposed into three following tensors $\boldsymbol{H}_{(1)} \in \mathbb{R}^{cH \times R}$, $\boldsymbol{H}_{(2)} \in \mathbb{R}^{cW \times R}$ and $\boldsymbol{H}_{(3)} \in \mathbb{R}^{D \times R}$.

The tensors $\boldsymbol{H}_{(1)}$ and $\boldsymbol{H}_{(2)}$ are retained and converted into 1-D vectors $\boldsymbol{V}_{(1)}$ and $\boldsymbol{V}_{(2)}$ as follows:

$$\boldsymbol{H}_{(1)} = \begin{bmatrix} H_{(1,1)} & \dots & H_{(1,R)} \\ & \vdots & \\ H_{(cH,1)} & \dots & H_{(ch,R)} \end{bmatrix} \rightarrow \boldsymbol{V}_{(1)} = \left[H_{(1,1)} \dots H_{(cH,1)} \dots H_{(1,R)} \dots H_{(cH,R)} \right] \tag{9}$$

```
procedure CP-DECOMPOSE($\boldsymbol{X}, R$)    ▷ $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, $R$-ranked
    for $n = 1 \to 3$ do
        Randomize $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$
        Normalize column vectors of $\mathbf{A}^{(n)}$
    end for
    repeat
        for $n = 1 \to 3$ do
            $Temp \leftarrow \boldsymbol{P}^{(3)} * \ldots * \boldsymbol{P}^{(n+1)} * \boldsymbol{P}^{(n-1)} * \ldots * \boldsymbol{P}^{(1)}$    ▷ $\boldsymbol{P}^{(i)} = \boldsymbol{A}^{(i)\top}\boldsymbol{A}^{(i)}$
            $\mathbf{A}^{(n)} \leftarrow \boldsymbol{X}^{(n)}(\mathbf{A}^{(3)} \odot \ldots \odot \mathbf{A}^{(n+1)} \odot \mathbf{A}^{(n-1)} \odot \ldots \odot \mathbf{A}^{(1)})\, Temp^{\dagger}$
            Normalize column vectors of $\mathbf{A}^{(n)}$ with $\boldsymbol{\lambda}$, update $\boldsymbol{\lambda}$
        end for
    until converged    ▷ Fit stops improve
    return $\boldsymbol{\lambda}, \boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \boldsymbol{A}^{(3)}$
end procedure
```

Fig. 3. ALP algorithm used on a third-order tensor

The values of the appended $\boldsymbol{V}_{(1)}$ and $\boldsymbol{V}_{(2)}$ will be the feature that represents the original image. The goal of the experiment is to evaluate the effectiveness of this type of feature against the unprocessed HOG descriptors. Beside the benefits gained from dimension reduction, it is suggested that this kind of feature will provide better accuracy in the image classification task.

3 Image Classification Experiment

We selected images of five classes which are 'Faces', 'Leopards', 'Motorbikes', 'airplanes' and 'car_side' from the Caltech-101 dataset [9]. There are 160 images selected per class. The HOG descriptors were created with both Dalal-Triggs and UoC/TTI variants. The classification performance of the HOG descriptors will be compared against the decomposed versions of each variant respectively.

Because the dimensions of sample images are not fixed, we utilized a conventional approach in image classification domain. The image descriptors will be fed into a K-means clustering bag-of-word (BoW) (VLFeat libray [10]) and the histograms of corresponding entries computed by accelerated Elkan optimization [10] of a test sample will be used as the features for SVM classifier. We used multi-class SVM which is an one-vesus-all linear SVM configuration, where the prediction for a sample $\mathbf{x}$ is based on the maximum probability among the SVMs: $\arg\max(\mathbf{w}^\top\mathbf{x} + \mathbf{b})$. Variable $\mathbf{w}$ is the weight vector and $\mathbf{b}$ is the bias of each SVM. The process is a 10-fold cross-validation and the classification performance is measured with three metrics: accuracy (true positive rate), sensitivity (positive rate) and specificity (negative rate).

3.1 Performance

We performed the experiment first with the R-Ranked of 1. Table 1 summarizes the result of the cross-validation. The columns with "CP" prefix contain the results when our method is applied with a HOG variant. One can see for both variants, the method provides features that have improved results over their non-decomposed version. There is also an interesting fact that our method yields identical classification result over the two cases.

Table 1. Classification performance comparison

	UoC/TTI	CP UoC/TTI	Dalal-Triggs	CP Dalal-Triggs
Accuracy	0.89	0.92	0.87	0.92
Sensitivity	0.89	0.92	0.87	0.92
Specificity	0.97	0.98	0.96	0.98

Figure 4 summarizes the impact of a chosen R-rank on the correct classification rate. The value 1 of R clearly yields the best result, as also due to the fact that this rank yields the fastest time to decompose with ALS. As R increases, the classification performance drops (down to the Dalal-Triggs's level when $R = 4$). Perhaps CP-decomposing an ordered set of descriptors with a value $R > 1$, the decomposed components does not follow that original order.

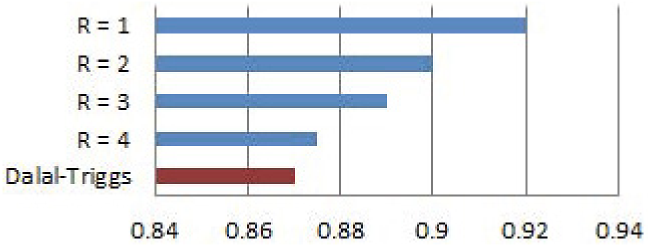
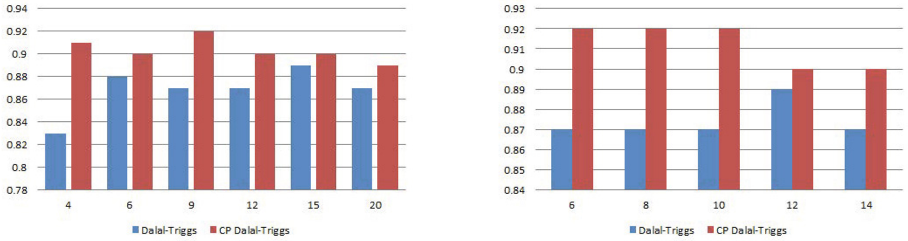


Fig. 4. Influence of R-Rank on classification accuracy

There is also an interest in the influence on accuracy from two HOG parameters: cell size (Figure 5a) and number of orientations (Figure 5b). As shown in both charts, the proposed method does consistently maintain an edge in terms of accuracy over the Dalal-Triggs variant.

4 Conclusion

This paper has revisited the popular HOG descriptor and combined it with tensor decomposition. This combination is not only to provide a more effective type of image descriptor but also showcasing the potential of this approach in computer vision field. The focus of this paper is on image classification problem, however there will be a lot more problems of this field could benefit from this.



(a) Number of orientations

(b) Cell size

Fig. 5. Effects of HOG parameters on overall accuracy

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (2005)
2. Khan, N., McCane, B., Wyvill, G.: Sift and surf performance evaluation against various image deformations on benchmark dataset. In: 2011 International Conference on Digital Image Computing Techniques and Applications (DICTA), pp. 501–506 (2011)
3. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
4. Jiang, J., Xiong, H.: Fast pedestrian detection based on hog-pca and gentle adaboost. In: 2012 International Conference on Computer Science Service System (CSSS), pp. 1819–1822 (2012)
5. Ke, Y., Sukthankar, R.: Pca-sift: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, pp. II–506–II–513 (2004)
6. Carroll, J., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. Psychometrika 35, 283–319 (1970)
7. Kiers, H.A.: Towards a standardized notation and terminology in multiway analysis. Journal of Chemometrics 14, 105–122 (2000)
8. Kolda, T., Bader, B.: Tensor decompositions and applications. SIAM Review 51, 455–500 (2009)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. 106, 59–70 (2007)
10. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008), `http://www.vlfeat.org/`

Phantom Elimination Based on Linear Stability and Local Intensity Disparity for Sonar Images[*]

Qiuyu Zhu and Yichun Li

School of Communication and Information Engineering
Shanghai University, Shanghai, P.R. China
zhuqiuyu@staff.shu.edu.cn, athrun_asuka@yahoo.com.cn

Abstract. The paper proposes a novel approach to the phantom elimination of sonar images based on image post-processing technique. Firstly, the images are transformed to the polar coordinate to form straight phantom. In the mapped images, the distribution of linear stability is further evaluated so that distance-direction positions of phantoms may be displayed by means of locating peak areas of linear stability. Then, the neighboring peak areas are combined to avoid mutual interferences. Lastly, the local intensity disparity of each peak area is calculated, which the inpainting strategies are taken to fulfill the inpainting work of phantom areas. The algorithm does not require mask images beforehand, and has good inpainting performance and a simple inpainting process.

Keywords: phantom elimination, linear stability, local intensity disparity, sonar image, image inpainting.

1 Introduction

Qualities of sonar images are related not only to the recognitions of underwater targets, but also to the establishment of underwater environmental surveillances in the future. Unfortunately, during the beamforming of sonar images, if echoes are too strong, an arc-shape bright line may be formed at the same distance of nearby beams, which is called a phantom academically. The most negative effect of phantoms is that they disturb the authenticity of sonar images and the distinction of underwater objects. Restraining sidelobes and ensuring the contribution of mainlobes to the distance-direction resolution by beam optimization is always a hot spot to deal with phantoms.

In order to eliminate image phantoms, guaranteeing the accuracy recognition of underwater objects, many researchers have made great efforts. Lei et al. [1] designed an FIR filter at the output of matched filter so that to minimize the ISL (Integrated Sidelobe Level), while keeping the mainlobe unchanged. Sun et al. [2] found an optimal modal in which beamforming problem could be formulated as a tractable convex second-order cone programming program, and the dimension of array weight vectors are decreased significantly by using the properties of spherical harmonics and Legendre polynomials.

[*] This work was supported by the Development Foundation of Shanghai Municipal Commission of Science and Technology (11dz1205902).

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 392–400, 2013.

In this way, the authors minimized the peaks of sidelobes while keeping the distortionless response in the look direction and maintained the mainlobe width. Wang et al. [3] considered the errors in sensor array characteristics together, and used worst-case performance optimization in mode space for circular arrays to minimize sidelobe beamforming. Liu et al. [4] introduced a mainlobe-to-sidelobe power ratio maximization constraint to the Capon beamforming to suppress the sidelobe for interference nulling. Huang et al.[5] proposed a novel blind beamforming algorithm based on the sparse beam pattern constraint for all possible interferences and sidelobes. Such algorithm overcame severe degradations under the conditions of unexpected interferences or high noise power. Hong et al. [6] derived the second order cone (SOC) formulation of a new adaptive beamformer which could be easily solved using the well-established interior point method. Berbakov et al. [7] proposed a DBF scheme with Sidelobe Control (SC) which maximized the beamforming gain in the direction of main BS while keeping the sidelobe levels in the directions of the unintended BSs below some prescribed thresholds. The proposed algorithm operated with partial CSI at the BSs and it merely required each BS to broadcast one bit of feedback to the sensor nodes. Sakhalin [8] established a generalized form of the weighting factor for the sidelobe reduction and restricted the apodization vector to a parametric representation through a discrete Fourier transform or discrete cosine transform which resulted in higher quality images with fewer artifacts and enhanced contrast properties. Belfiori et al.[9] retrieved the phase modes pattern for the reference array and then applied a conventional tapering for the sidelobe suppression of the obtained virtual uniform linear array.

Due to the limitations of sonar equipment, sonar echoes inevitably contain sidelobe ingredients, thus, the sonar images also inevitably contain circular phantom. The paper proposed a novel solution to the phantom elimination for sonar images based on image post-processing technique. The rest of the paper is organized as follows. In the section 2 the phantom elimination method based on linear stability and local intensity disparity are introduced. Section 3 exhibits experimental results and discussion. Section 4 concludes the paper.

2 The Algorithm of Phantom Elimination Based on Linear Stability and Local Intensity Disparity

The representation of the sonar image is a sector image. In image, phantoms, a kind of image artifacts, exist in the form of circular arcs crossing the picture. A sonar image with several phantoms is shown in Fig. 1.

It has been demonstrated that phantoms are caused by interferences of sidelobes. However, due to the limitations of sonar equipment, sonar echoes inevitably contain sidelobe ingredients. It is widely known that Hough circle detection is a good method to seek arc and circle targets[10]. But due to the inhomogeneity of phantom areas, besides noises and large objects around the phantoms, these factors increase image complexity, making Hough circle detection suffer time-costing and low accuracy. From another perspective, the fact that phantoms have no information to transfer the problem into handling fragmentary images, image inpainting[11] will be a good technical solution. Thus the paper puts forward a novel image post-processing method to deal with phantoms. According to the image and phantom characteristics, the algorithm includes the following two main steps: (1) phantoms positioning based on linear stability (2) image inpainting based on local intensity disparity.

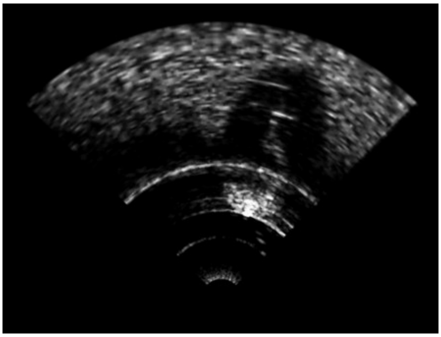
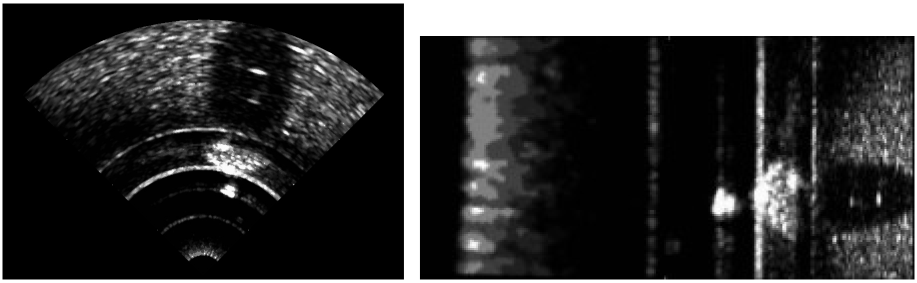


Fig. 1. A sonar image with phantoms

2.1 Phantoms Positioning Base on Linear Stability

Polar Projection. Because sonar images studied are sector images, with already known sector radius R', circle center O, angle A, rectangle images with the size of R'×A could be acquired by mapping images from Cartesian coordinates to polar coordinates. Doing that, processing the images by column and row is more efficient than by radius and angle. The procedure of polar projection shows in Fig.2.



(a) An original sector sonar image (b) The rectangle mapping image after polar projection

Fig. 2. Polar projection of a sonar image

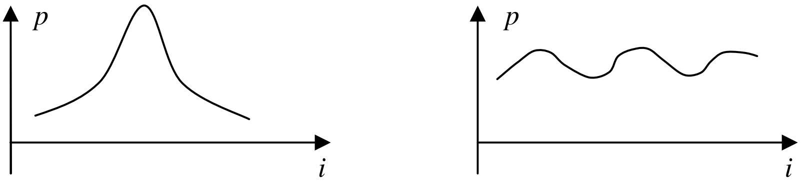
Locating Peak Areas of Linear Stability. After projection, the arc phantoms appear in the form of bars. Observed them by column, each column of them is nearly a straight line. Fig.3 gives the column mapping images of an object and a phantom separately from Fig.2(b). The images in Fig.3 are properly zoomed in for better watching.

It can be found out from Fig.3 that the waveform of object intensities tends to unimodal distribution while the one of phantom intensities is slight-shaking wave distribution, as shown in Fig.4.

Therefore, seeking phantoms turn into the question of distinguishing between unimodal and wave distribution. A classical approach to object classification in pattern recognition is clustering. Usually nice results mean small distances in clusters and large distances between clusters. It is suitable to point out the phantoms by means of using the distances in clusters of the waveforms of intensities.

(a) The column mapping image of an object (b) The column mapping image of a phantom

Fig. 3. Column mapping images of an object and a phantom from the rectangle mapping image



(a) Unimodal distribution (object) (b) Wave distribution (phantom)

Fig. 4. The waveforms of object and phantom intensities in column mapping images

Thus linear stability is proposed in this paper, the function can be expressed as:

$$D_x = \frac{K \bullet Avg_x}{Var_x}, x = 1 \ldots R' \tag{1}$$

where Avg$_x$ is the intensity mean of x column of the mapping image, Var$_x$ is the intensity distribution variance of x column, K is a constant. Further linear stability distribution could be obtained like Fig.5.

Local peak searching is performed towards linear stability distribution, whose searching window size is W. Then the center of the window is set on each peak (labeled as P). If attenuation of linear stability values of the area around P reach the preset peak acutance threshold (labeled as Pt), the area is called a peak area of linear stability. Position information of the area is saved. Each peak area corresponds to a phantom bar.

One point should be paid attention to is that sometimes two peak areas may be too close that the bigger could cover the smaller. It leads to phantom missing. To deal with this, whenever a peak area is positioned, the linear stability values among the peak area should be reduced so that the possibly-existing smaller areas around the peak area could be highlighted.

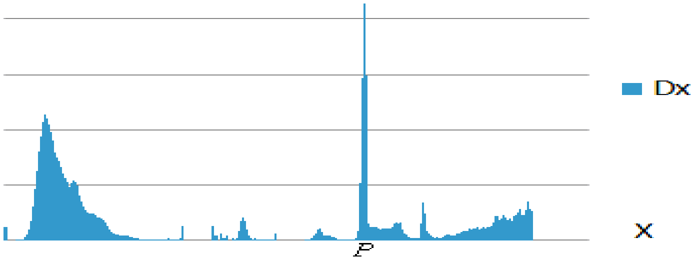


Fig. 5. Linear stability distribution of a sonar image

2.2 Image Inpainting Based on Local Intensity Disparity

After all the peak areas of linear stability have been determined, phantoms could be positioned in distance direction. But considering the fact that phantoms maybe cross objects in sonar images, there are overlapped regions between phantoms and objects. In other word, adjacent connected regions of such phantoms belong to the objects. Although phantoms have no information, covering these overlapped regions with zero in inpainting course is rather inappropriate. Thus, belongingness of adjacent connected regions of phantoms should be decided, which guides the inpainting work of phantoms.

Combination of Neighboring Peak Areas of Linear Stability. Because belongingness of adjacent connected regions of phantoms involves a small field around each phantom, two neighboring phantoms could interfere with each other, for the nearby phantom may be judged as an object by mistake. Misjudgment affects the choices of inpainting strategies of phantoms. In order to avoid this event, if two peak areas of linear stability are closer than preset spacing threshold Gap, the areas will be merged into a new large peak area to ensure the fitness of subsequent calculations, which realizes combination of neighboring phantoms.

Local Intensity Disparity. Observing a nearby field of each phantom located by corresponding peak area of linear stability by angle, the belongingness of phantom region has three situations: (1) both left and right adjacent connected regions of the phantom belong to background, whose local intensity distribution is unimodal like Fig.4(a). (2) Both regions belong to objects. Its local intensity distribution is a waveform like Fig.4(b). (3) One side adjacent regions belong to objects while the others belong to background, whose local intensity distribution is shown as Fig.6.

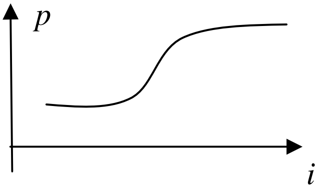


Fig. 6. The local intensity distribution under the condition that one side adjacent connected regions of phantom belong to objects meanwhile the other side belong to backgrounds

Furthermore, local intensity disparity is proposed in this paper to indicate these situations mentioned above. The equation is:

$$\begin{cases} DY_L(L-SH,j)=\dfrac{T(L-SH,j)}{T(P,j)} \\ DY_R(R-SH,j)=\dfrac{T(R-SH,j)}{T(P,j)} \end{cases}, j=1\ldots A \tag{2}$$

where T(*P*,*j*) stands for the intensity of the pixel at (*P*,*j*) in mapping images (*P* is the location of the peak of a peak area) while T(*L*-*SH*,*j*) is the intensity at (*L*-*SH*,*j*) and T(*R*-*SH*,*j*) is the intensity at (*R*-*SH*,*j*) (*L*,*R* are the locations of left and right endpoints of a peak area). *SH* is a offset. Then an important local intensity disparity threshold DT needs to be set. Then belongingness of adjacent connected regions of phantoms can be gained. With the help of the belongingness and choices of inpainting strategies:

$$T(m,n)=\begin{cases} 0,(DY_L(L-SH,n)<DT)\cap(DY_R(R-SH,n)<DT) \\ T(m,n),(DY_L(L-SH,n)>DT)\cup(DY_R(R-SH,n)>DT) \end{cases}, m=L\ldots R, n=1\ldots A \tag{3}$$

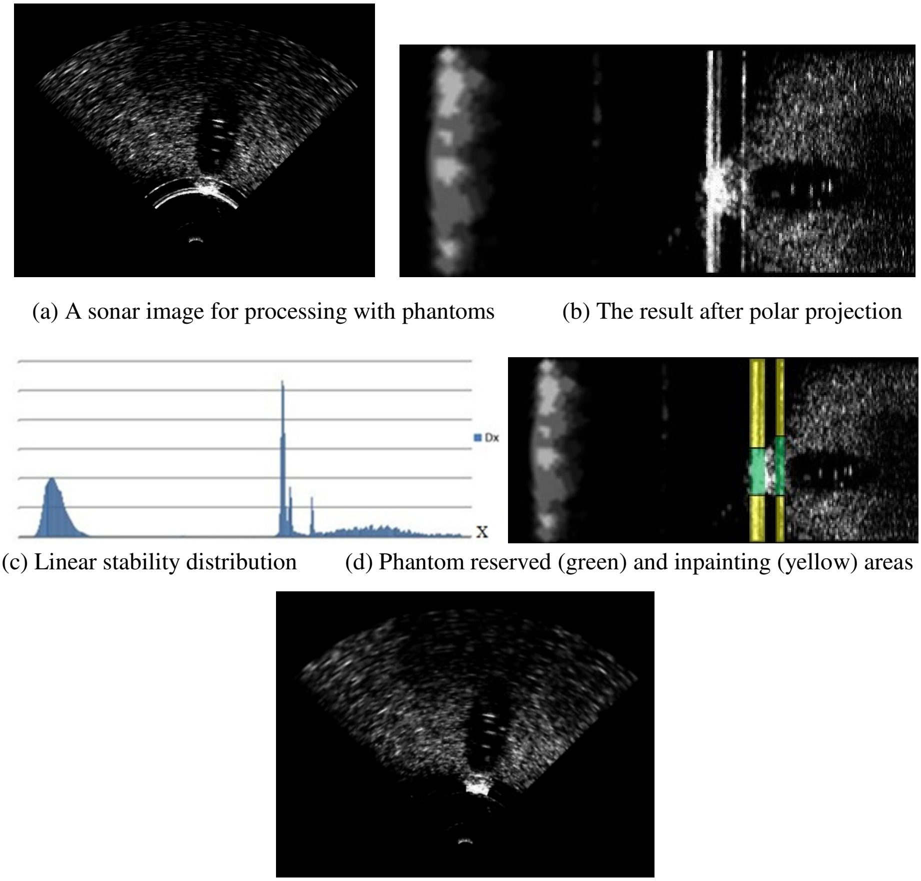
where *T*(*m*,*n*) is the intensities of the pixels inside phantoms, phantom elimination comes true. When at least one side adjacent connected regions of a phantom belong to objects in a row, all the pixels of the row are to be reserved for protecting object information, which form a phantom reserved area. Likewise, when at least one side adjacent connected regions are background in a row, all the pixels of the row are to be padded with zero-intensity pixels according to the fact that phantoms have no information. Also, a phantom zero-intensity inpainting area emerges.

3 Experimental Results and Discussions

To verify the effectiveness of the algorithm, functions expected are realized with C++ development language in the Microsoft Visual 2008 C++ environment, combined with open source computer vision library OpenCV. A sonar image for processing is shown as Fig. 7(a). In the central and lower part of Fig. 7(a), a target with high brightness is passed through by three phantoms. The recognition of the exact size of the target worsens. Firstly polar projection acts on the image, the result is shown in Fig. 7(b). Then linear stability of the mapping image is calculated by column, getting the linear stability distribution as Fig. 7(c). The three sharp peak areas represent the three phantoms in

Fig. 7(c). Peak acutance threshold Pt is set to 0.45. Behind this, all peak areas can be located by local peak searching results and the elaborate Pt. Over combination of neighboring peak areas of linear stability and computation of local intensity disparity, phantom reserved areas (green areas) and phantom zero-intensity inpainting areas (yellow areas) could be obtained as Fig. 7(d). Here in order to balance the sensitivity of targets and backgrounds, after a lot of testing, local intensity disparity threshold DT is finally set to 0.5. With respective inpainting strategies, the sonar image after phantom elimination is shown in Fig. 7(e). The sonar image information has been enhanced by means of eliminating the phantoms effectively.

To assess phantom elimination based on linear stability and local intensity disparity under challenging conditions, quantitative comparison among the approach in this paper and two widely used image inpainting methods, i. e. FMM [12] and Navier-Stokes [13], is made. Fig. 8 shows the results of the last two algorithms. The inpainting



(a) A sonar image for processing with phantoms

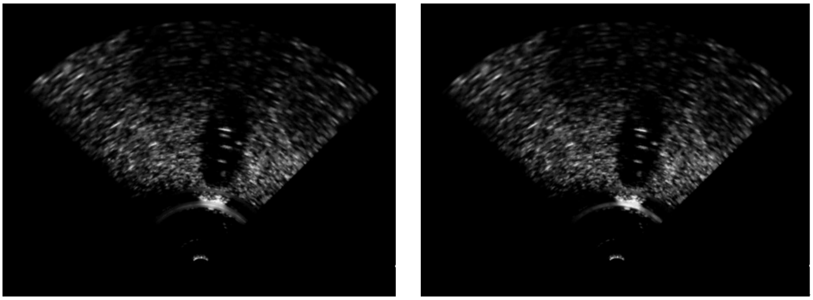
(b) The result after polar projection

(c) Linear stability distribution

(d) Phantom reserved (green) and inpainting (yellow) areas

(e) The result after phantom elimination

Fig. 7. Using linear stability and local intensity disparity to realize phantom elimination



(a) The result of FMM method

(b) The result of Navier-Stokes method

Fig. 8. Phantom elimination results of other algorithms

results of them have residual phantoms, greatly owing to the existence of neighboring phantoms. In contrast, the method proposed in this paper has better ability of restraint and elimination of phantoms on the basis of the result in Fig. 7(e).

4 Conclusions

The paper puts forward an approach to eliminate phantoms for sonar images based on image post-processing. Peak areas of linear stability reveal distance-direction positions of phantoms. Local intensity disparities of each peak areas cooperated with choices of inpainting strategies guide the inpainting work of phantoms. Experimental results show that the algorithm has good effectiveness, strong robustness, and lower algorithm complexity.

References

1. Lei, B., Yang, K.D., Wang, Y.: Optimal sidelobe reduction of matched filter for bistatic sonar. In: 2012 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM), Zhangjiajie, China, pp. 469–472 (March 2012)
2. Sun, H.H., et al.: Robust minimum sidelobe beamforming for spherical microphone arrays. IEEE Transactions on Audio, Speech, and Language Processing 19(4), 1045–1051 (2011)
3. Wang, Y., et al.: Robust minimum sidelobe beamforming in mode space for circular arrays. In: 2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC), Hong Kong, pp. 300–304 (2012)
4. Liu, Y., Wan, Q.: Sidelobe suppression for robust capon beamforming with mainlobe-to-sidelobe power ratio maximization. IEEE Antennas and Wireless Propagation Letters 11, 1218–1221 (2012)
5. Huang, J.Y.Y., Wang, P., Wan, Q.: Sidelobe suppression for blind adaptive beamforming with sparse constraint. IEEE Communications Letters 15(3), 343–345 (2011)
6. Hong, Z.Q., et al.: Adaptive beamforming for MIMO radar with sidelobe control based on second order cone programming. In: 2012 International Conference on Information Science and Technology (ICIST), Wuhan, China, pp. 384–388 (2012)
7. Berbakov, L., Anton-Haro, C., Matamoros, J.: Distributed beamforming with sidelobe control using one bit of feedback. In: 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), Budapest, Hungary, pp. 1–5 (2011)

8. Sakhaei, S.M.: Optimum beamforming for sidelobe reduction in ultrasound imaging. IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control 59(4), 799–805 (2012)
9. Belfiori, F., et al.: Side-lobe suppression techniques for a uniform circular array. In: 2010 European Radar Conference (EuRAD), Paris, France, pp. 113–116 (2010)
10. Nosrati, M., Karimi, R.: Detection of circular shapes from impulse noisy images using median and laplacian filter and circular hough transform. In: 2011 8th International Conference on Electrical Engineering Computing Science and Automatic Control (CCE), Merida, Mexico, pp. 1–5 (2011)
11. Lorenzi, L., Melgani, F., Mercier, G.: Inpainting Strategies for Reconstruction of Missing Data in VHR Images. IEEE Geoscience and Remote Sensing Letters 8(5), 914–918 (2011)
12. Telea, A.: An image inpainting technique based on the fast marching method. Journal of Graphics Tools 9(1), 25–36 (2004)
13. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier-stokes: fluid dynamics, and image and video inpainting. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 355–362 (2001)

Composite Color Invariant Feature H' Applied to Image Matching

Keisuke Kameyama[1] and Wataru Matsumoto[2]

[1] Faculty of Engineering, Information and Systems, University of Tsukuba
Keisuke.Kameyama@cs.tsukuba.ac.jp
[2] Graduate School of Systems and Information Engineering, University of Tsukuba,
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Abstract. Object color is one of the most important feature in camera-based object matching. Color invariants are features based on the models of color observation that tends to be constant under varying conditions of illumination and surface. In this work, we analyze the estimation process of the color invariants by Geusebroek et al. from RGB images, and propose a novel invariant feature H' based on the elementary invariants to meet the circular continuity residing in the mapping between colors and their invariants. The use of the proposed invariant in combination with luminance, contributes to improve the retrieval performances of partial object image matching under varying illumination conditions.

Keywords: Color Invariants, Image Matching, CBIR, SIFT.

1 Introduction

In image matching, color is an important descriptor for finding the matching object. However, the apparent color of objects can change drastically according to illumination, surface type and observation condition.

Geusebroek et al. introduced a group of color invariants that are specific measures of the object surface color, that can be estimated under certain illumination conditions [3]. The invariants are based on Kubelka-Munk theory of surface observation and the Gaussian color model. The invariants derived in the work has been successfully used as image descriptors in place of grayscale for solving keypoint correspondence [1]. Among the various color-based image descriptors, it has been reported to have robust results for use in object and scene matching [10].

In this work, we analyze the estimation process of the invariants from RGB colors, and propose a novel color invariant feature H' defined as a composite function of two basic invariants. The novel invariant intends to improve the selectivity of the corresponding object colors so that identical objects under different illumination will be precisely matched.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 401–408, 2013.

2 Color Invariants

In [3], Geusebroek et al. introduced a set of illumination invariant color features (color invariants) from a surface reflection model used in Kubelka-Munch theory [6] with assumptions similar to those used in the Dichromatic Reflection Model by Shafer [9]. Among the color invariants, we will focus on two invariants named H and C in [3]. The two can be estimated from the RGB components of digital images assuming the Gaussian color model that approximates the Hering basis [5] corresponding to the chromatic human vision [3].

The first invariant H for white illumination is defined as, and can be estimated as

$$H = \frac{\frac{\partial E}{\partial \lambda}}{\frac{\partial^2 E}{\partial \lambda^2}} = \frac{0.3R + 0.04G - 0.35B}{0.34R - 0.6G + 0.17B}. \tag{1}$$

Here, E is the spectrum of surface reflection which is a function of wavelength λ, and R,G,B are the color components. This H is invariant to shadows, highlights and change of illumination intensity.

The second invariant C for white illumination on a matte surface is defined as, and is estimated as

$$C = \frac{1}{E(\lambda)} \frac{\partial E}{\partial \lambda} = \frac{0.3R + 0.04G - 0.35B}{0.06R + 0.63G + 0.27B}. \tag{2}$$

This C is invariant to shadows and changes of illumination intensity.

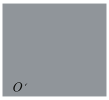
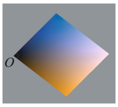
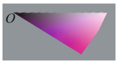
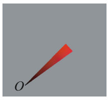
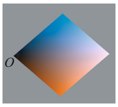
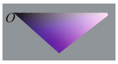
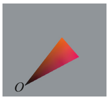
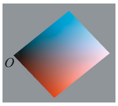
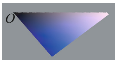
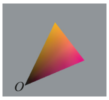
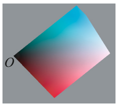
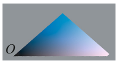
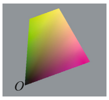
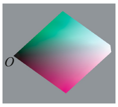
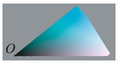
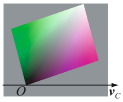
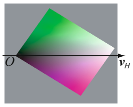
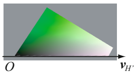
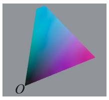
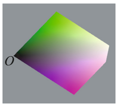
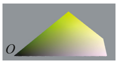
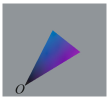
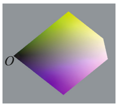
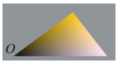
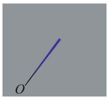
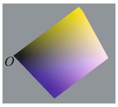
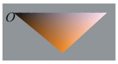
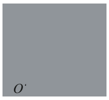
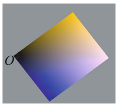
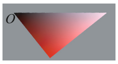
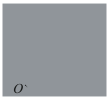
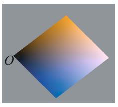
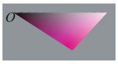
A comprehensive set of invariants under different illumination and surface conditions, and those derived using higher-order derivatives of $E(\lambda)$ have been mentioned in [3]. In this work, however, we will only discuss the use of fundamental invariants H and C in color invariant-based matching of objects.

Invariants H and C have the following properties [7].

1. Set of points (colors) in the RGB space having the same invariant values will be a 2-dimensional subspace (plane), namely the equal-H plane $P_h(H)$ and equal-C plane $P_c(C)$, respectively.
2. Planes $P_h(0.1111)$ and $P_c(-0.0104)$ include line $R = G = B$ for achromatic colors.
3. $P_h(0) = P_c(0)$.
4. The planes will rotate π radian as the invariants change from $-\infty$ to ∞. Therefore, $P_h(-\infty) = P_h(\infty)$ and $P_c(-\infty) = P_c(\infty)$.
5. The rotating axes of $P_h(H)$ and $P_c(C)$ are $\boldsymbol{v}_H = (0.6193, 0.5181, 0.5900)$ and $\boldsymbol{v}_C = (0.7361, -0.3246, 0.5939)$, respectively.
6. Plane $P_h(H)$ passes through the RGB cube (where colors are assigned in a framebuffer) for any $H \in (-\infty, \infty)$. In contrast, plane $P_c(C)$ passes through the RGB cube only for C in a certain interval.

Since the values of H and C are spanned over $(-\infty, \infty)$, often it is more convenient to use $\theta_H = \arctan(H) \in (-\frac{\pi}{2}, \frac{\pi}{2})$ and $\theta_C = \arctan(C) \in (-\frac{\pi}{2}, \frac{\pi}{2})$ instead. This convention will be used in the following.

Table 1. Set of colors having equal invariants C, H and H' for several values of invariants in arctan convention in $(-\pi/2,\ \pi/2)$

$\theta_C/\theta_H/\theta_{H'}$	$P_c(\theta_C)$	$P_h(\theta_H)$	$P_{h'}(\theta_H)$
$(-\pi/2)$ -1.5			
-1.2			
-0.9			
-0.6			
-0.3			
0			
0.3			
0.6			
0.9			
1.2			
1.5 $(\pi/2)$			

3 Composite Color Invariant H'

In Table 1, colors on equal-C plane P_c and equal-H plane P_h are shown for different values of invariants θ_C and θ_H. The colored regions show the cross sections of the RGB cube as the planes rotate. In each map, the origin of the RGB space is shown as O. Vectors $\boldsymbol{v}_C$ and $\boldsymbol{v}_H$ are the axes of rotation of P_h and P_c, respectively. It is found that colors of different hues (blue-orange, cyan-red, green-purple, etc) coexist on P_h with a band of achromatic color in between. From an application point of view, object regions having different colors being attributed to same feature value can lead to false matching.

In order to avoid this issue, a new invariant that separates the colors at the achromatic color band is introduced. On characterizing the achromatic colors, we chose to use the value of C, at which it takes the value of $C = C_0 \approx -0.0105$.

The angular expression of the new invariant H' which addresses this issue is defined as,

$$\theta_{H'} = \begin{cases} \frac{1}{2}\theta_H - \frac{\pi}{2} & (\theta_C \geq \theta_{C0}, \theta_H \geq 0) \\ \frac{1}{2}\theta_H & (\theta_C \geq \theta_{C0}, \theta_H < 0) \\ \frac{1}{2}\theta_H & (\theta_C < \theta_{C0}, \theta_H \geq 0) \\ \frac{1}{2}\theta_H + \frac{\pi}{2} & (\theta_C < \theta_{C0}, \theta_H < 0) \end{cases} \tag{3}$$

The rightmost column of Table 1 shows the sets of colors that have identical H' values. The H' being a function of H and C, it is invariant to shadow and change of illumination intensity. Most importantly, it has a color correspondence similar to hue and solves the issue of H pointed out above.

4 Image Matching Experiment

The color invariants H, C, H' and their combinations will be used as features for local descriptors in object matching under varying illumination conditions. For comparison, luminance and hue were also used. Matching will be based on finding the correspondence of keypoints in the two images and finding the corresponding affine-transformed regions.

In the database, images of objects taken under different illumination conditions and observation angles were kept. The query images are partial images of the objects also with variations in scaling, rotation and illumination conditions. Here, the difference in the retrieval performance according to the employed image feature set was evaluated.

Local Descriptor. Scale-Invariant Feature Transform (SIFT) [8] using illumination invariant features known as Colored SIFT : (CSIFT) [1] was used as local descriptors.

Distance Measure. In evaluating the feature discrepancy between two values θ_1 and θ_2 ($-\pi/2 \leq \theta_2 \leq \theta_1 < \pi/2$) of an invariant or hue, a cyclic distance measure

$$d_{\text{cyclic}}(\theta_1, \theta_2) = \min((\theta_1 - \theta_2), \{(\theta_2 + \pi) - \theta_1\}) \tag{4}$$

was used.

Correspondence of Features and Regions. Matching of CSIFT descriptor was based on the nearest neighbor search in the feature space. A union of corresponding point pairs were used in resolving the matching regions in the two images. The pairs were further interpreted as matched regions by finding the homography between the regions by RANdom SAmple Consensus (RANSAC)[2]. The goodness of match for the two regions were evaluated using the normalized cross-correlation (NCC) of the corresponding regions.

Evaluation. For evaluating the correspondences of two images, the F-measure defined as

$$J_{\mathrm{F}} = \frac{2N_{\mathrm{TP}}}{(N_{\mathrm{TP}} + N_{\mathrm{FN}}) + (N_{\mathrm{TP}} + N_{\mathrm{FP}})} \tag{5}$$

is used. Here, N_{TP}, N_{FP} and N_{FN} denote the numbers of image matches judged as true positive (TP), false positive (FP) and false negative (FN), respectively.

Image Set. The images used in the experiments were 50 object images selected from the Amsterdam Library of Object Images (ALOI) dataset[4]. ALOI is a collection of common objects having various color and texture, observed from various angles under various camera and illumination positions. It also includes images observed under different illuminations. However, only images illuminated by a tungsten halogen lamp with near-flat continuous spectrum have been used for this work. All images have 384×288 pixel dimensionality, with pixels having 24 bit depth of RGB color. The images were compiled into two sets focusing on different observation conditions.

- **Set Illum-Angle**
  Images taken under different illumination positions and camera angles. Includes 50 (objects) × 8 (illuminations) × 3 (angles) = 1200 images.
- **Set Rotation**
  Images of objects rotated horizontally $-45°$, $-30°$, $-15°$, $0°$, $15°$, $30°$ and $45°$ against the camera. Includes 50 (objects) × 7 (rotation angles) = 350 images. Examples of the images for an identical object is shown in Fig. 1(a).

Query Image. The partial query images were cut out from several database images. Altogether, a total of 73 query images were used. For each object, three query images were prepared as follows.

- **Query A** : Partial image of an object enlarged by 200% : 26 images (Fig. 1(c)).
- **Query B** : Partial image in Query A rotated 45°, in its original size : 26 images (Fig. 1(d)).
- **Query C** : Same portion as Query A from a different image which is significantly darker : 21 images (Fig. 1(e)).

Image feature. Features used for keypoint matching in CSIFT and region matching were luminance L, invariants H, C, H', Hue, luminance-invariant pairs $\{L, H\}$, $\{L, C\}$, $\{L, H'\}$ and $\{L, Hue\}$. Upon retrieval, all matched regions in the database images having positive NCC ($R_N > 0$) to the query were included in the retrievals.

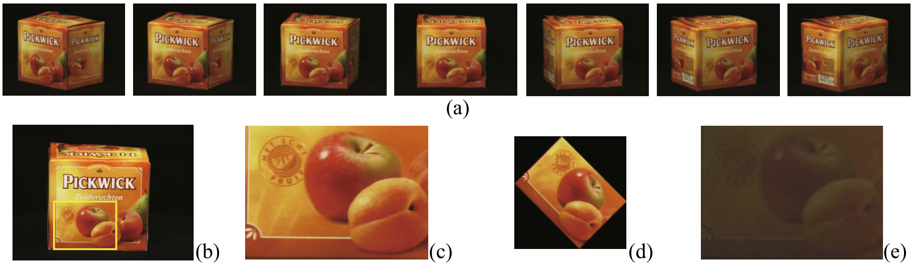


Fig. 1. (a) Examples of Rotation images. Query images : (b) Original object image. (c) Query A from the framed area of (b), (d) Query B (rotated Query A), and (e) Query C (darker Query A).

Correct Retrieval. For each query image, a set of images meeting either of the following criteria were selected as the correct retrieval.

- The object in the database image was identical with the one from which the query image was selected (having the same object ID).
- The image included the query image regardless of the object ID (e.g. having the same logo).

However, matching to objects of different colors were judged to be false even if the patterns or texts were identical.

Results. F-measure for different types of queries and combinations of features are shown in Fig. 2.

For the scaled (Query A) and rotated (Query B) queries, luminance L achieved better correspondence than the invariants, probably because there were less differences in the illumination intensity between the query and database images.

When dark queries (Query C) were used, the advantage of using invariants C and H' became obvious. In contrast, the performance dropped significantly for L and H. Improvement from H to H' shows that confining the equal-H' color set to similar hue colors was a plus.

Except for Query C in Illum-Angle image set where joint use of L with C or H' had a slightly negative contribution, jointly using L and other invariants contributed to improve performances.

Performances when Hue was used as the descriptor was generally low, which was unexpected. Although further investigation is due, it can be considered to be due to its high variance against different illumination conditions in the dataset.

5 Discussion

The reason for the superior performances for color invariants C and H' may be explained by the actual invariance of the descriptors under varying illumination conditions. In Fig. 3, the variance of L, H, C, H' and Hue are compared through

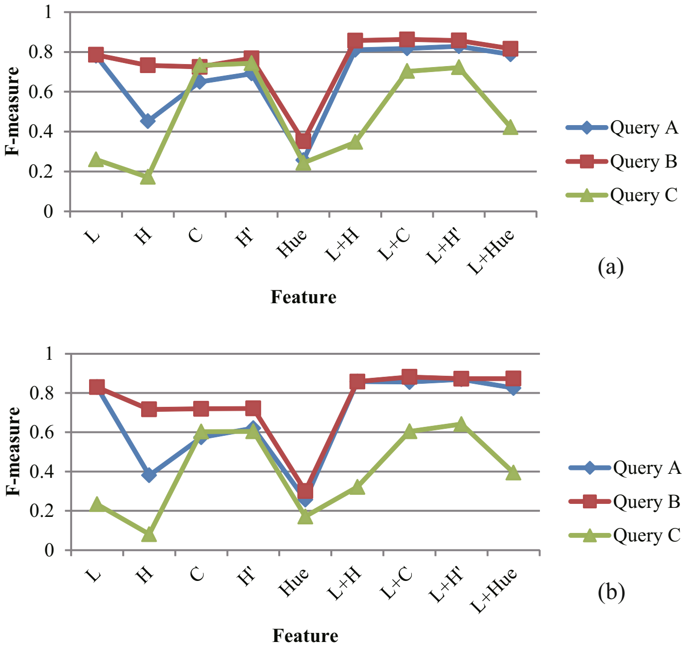


Fig. 2. Average F-value of retrieval for images in (a) Illum-Angle and (b) Rotation

the 8 different lighting conditions for selected positions of 7 objects included in the image set Illum-Angle. The variances are calculated for each descriptor after normalizing the ranges of the descriptors to unity.

It is clear that the stability of the descriptor is directly reflected in the image matching performances under varying illumination conditions in Fig. 2. The Hue is supposed to be invariant under an ideal white illumination, however, results in Fig. 3 show that this is not the case for the images used in the experiment. Further investigation and tuning of invariant measures for specific observation conditions should be beneficial for real-world implementation of illumination invariant object recognition.

6 Conclusion

In this work, a composite color invariant H' as a function of color invariants C and H was proposed to improve the selectivity of object color under varying lighting conditions. Luminance, hue and different invariants were compared as image descriptors in matching of partial object images by way of keypoint matching using CSIFT and region matching based on RANSAC. Joint use of L and C or H' gave the most stable performances for different query conditions, and the merits of using the invariants in image matching under varying illumination conditions were shown.

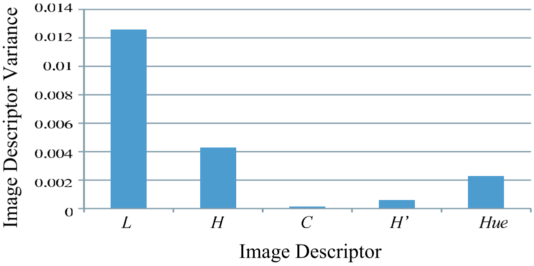


Fig. 3. Variances of L, H, C, H' and Hue under different illumination angles

References

1. Abdel-Hakim, A.E., Farag, A.A.: CSIFT: A SIFT descriptor with color invariant characteristics. In: Proceedings of IEEE CVPR 2006, pp. 1978–1983 (2006)
2. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of ACM 6, 381–395 (1981)
3. Geusebroek, J.M., van den Boomgaard, R., Smeulders, A.W.M., Geerts, H.: Color Invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(12), 1582–1595 (2001)
4. Geusebroek, J.M., Burghouts, G.J., Smeulders, A.W.M.: The Amsterdam library of object images. International Journal Computer Vision 61(1), 103–112 (2005)
5. Hering, E.: Outlines of a theory of the light sense. Harvard University Press (1964)
6. Judd, D.B., Wyszecki, G.: Color in Business, Science and Industry, 2nd edn. Wiley (1963)
7. Kobayashi, M., Kameyama, K.: A Composite Illumination Invariant Color Feature and its Application to Partial Image Matching. IEICE Transactions on Information and Systems E95-D(10), 2522–2532 (2012)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110 (2004)
9. Shafer, S.A.: Using color to separate reflection components. Color Research and Application 10(4), 210–218 (1985)
10. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1338–1350 (2010)

GPU-Based Real-Time Pedestrian Detection and Tracking Using Equi-Height Mosaicking Image

Min Woo Park and Soon Ki Jung

The School of Computer Science and Engineering, Kyungpook National University,
80 Daehak-ro, Buk-gu, Daegu 702-701, Repulic of Korea
mwpark@vr.knu.ac.kr, skjung@knu.ac.kr

Abstract. In this paper, we present a GPU-based real-time pedestrian detection and tracking system using a novel image representation called the equi-height mosaicking image [1]. This representation improves the processing time of the existing acceleration approach to pedestrian detection without decreasing accuracy. In equi-height mosaicking image generation, we first detect the horizon and crop a set of image strips from the road at uniform distance intervals. The height of each image strip is computed by projecting the predefined average height of a pedestrian at that distance onto the image plane. Then, all cropped images are resized to a uniform height and concatenated into a panorama image. Next, we detect the pedestrians on an equi-height mosaicking image using 1D based SVM classification. The SVM classifier is trained by an image dataset generated from various heights of pedestrians. After finishing this detection, we track the detected pedestrian in the previous frame. We performed the matching process in the neighbor block area of the equi-height mosaicking image to restrict the computation region. The detected or tracked results mapped onto the original image and grouped into multiple, overlapping regions.

Keywords: Pedestrian, Detect, GPU, SVM, HOG, Equi-Height Image.

1 Introduction

Today many people are injured or killed in traffic accidents. In order to prevent this, much recent research has focused on developing active safety systems such as blind spot warning (BSW) systems or forward collision warning systems (FCWS). In these safety systems, an intelligent vehicle protects both driver and pedestrians from the driver's mistakes. Much research has recently focused on a vision-based safety system that prevents collisions with pedestrians. The vision-based system is mainly used a general-purpose device in an intelligent vehicle due to its low cost. In order to increase the processing speed, many hardware acceleration algorithms have been designed, for example, GPU (graphic processing unit) or FPGA (field programmable gate array) algorithms.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 409–416, 2013.

Dalal and Triggs [2] have invented a pedestrian detection method that uses Histograms of Oriented Gradients (HOG) as these show a high accuracy for object detection. However, they consume a lot of processing time. Therefore, recent studies have focused on improving this. Prisacariu and Reid [3] have implemented an acceleration approach using a hardware accelerator. It is faster than a CPU-based non-acceleration approach. Rodrigo Beneson et al. [4] have created a fast GPU-based pedestrian detection method using the stixels computed by a stereo camera. However, the processing time of both these systems depends on the number of image pyramid levels for the pedestrians scaling. When we increase the number of image pyramid levels to obtain higher accuracy or to detect smaller pedestrians, the increased processing regions rapidly slow execution. Therefore, we propose an efficient image representation that will increase the execution speed by decreasing the number of computation regions. It is also an efficient tracking method.

The remainder of our paper is organized as follows. In Section 2, we describe the generation of equi-height mosaicking images for fast processing. The pedestrian detection and tracking method using equi-height mosaicking images is described in Section 3, along with experiments in Section 4. Finally, concluding remarks are presented in Section 5.

2 Equi-Height Mosaicking Image Generation

In this section, we introduce the equi-height mosaicking image, which was originally proposed in our previous work [1]. The entire equi-height mosaicking image generation process is performed on GPU device memory in order to ensure fast processing.

2.1 Calibration and Horizon Detection

First, we perform calibration and horizon detection in pre-processing step. Calibration means the compensation for lens distortion and the skew of the image. This calibration is performed using camera calibration parameters and an estimated skew angle [5,6].

After calibration is complete, we estimate the vanishing point on the image to detect the horizon. The vanishing point is identified by observing road lane based clustering using the Probabilistic Hough Transform [7] and MSAC [8,9]. Next, we estimate the vanishing point using the least square solution from the clustered lines.

2.2 Hypothesis Position Sampling and Height Estimation

To generate equi-height images, the proposed system performs hypothesis position sampling. This means the extraction of y-coordinates on the image at uniform distance intervals from the camera position as shown in Fig. 1. In order to perform hypothesis position sampling at uniform distance intervals, we extract the hypothesis positions from a bird's eye view image [10]. In this case, we set the uniform distance intervals at 62.5cm in three dimensional space because the average walking step of women and men is respectively 66.04cm and 78.74cm [11]. Therefore, we can detect the pedestrian within 25m by using 40 sampling levels at intervals of 62.5cm. After the position sampling is finished, we inverse warp the sampling coordinates from a bird's eye view image to the original image. Next, we generate equi-height images using the inverse warped sampling coordinates because any pedestrian may be located at the sampling positions. We estimate the height of image strips using the following equation,

$$R_h/(R_p - l_h) = O_h/(O_p - l_h), \tag{1}$$

where R_h and R_p are reference height and position, respectively. O_h and O_p are, respectively, the height and position of the pedestrian at the sampling positions. l_h is the horizon. Fig. 1 illustrates the equi-height image generation of sampled positions in 3D space. S_1, S_2 and S_3 are sampling positions. h_p and h_c are, respectively, the average height of the pedestrian and the height of the camera.

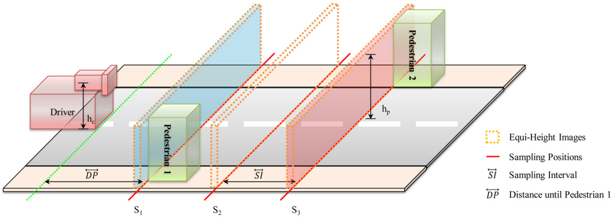

Driver
h_c
Pedestrian 1
Pedestrian 2
h_p
$\overleftrightarrow{DP}$
$\overleftrightarrow{SI}$
S_1
S_2
S_3
Equi-Height Images
Sampling Positions
$\overleftrightarrow{SI}$ Sampling Interval
$\overleftrightarrow{DP}$ Distance until Pedestrian 1

Fig. 1. Equi-height image generation on sampled positions in 3D space [1]

2.3 Equi-Height Image Generation and Mosaicking

As in our previous research [1], we generate equi-height images at each sampling position and concatenate them into a panorama image for fast detection and tracking. The term equi-height image means the cropped image strip using the average height of pedestrians on the image. Fig. 2 illustrates equi-height image generation on a 2D image.

Finally, we concatenate these images into a pushbroom or panorama image. In order to do so, we resize the equi-height images at a fixed image height equal to the size of training data. After resizing, we concatenate them into a long image. The generated equi-height mosaicking image is used as input for pedestrian detection and tracking. Fig. 4 shows the generated equi-height mosaicking image.

3 Pedestrian Detection and Tracking

After the equi-height mosaicking image is successfully generated, we perform a detection and tracking process to find the pedestrian's location. Therefore, we

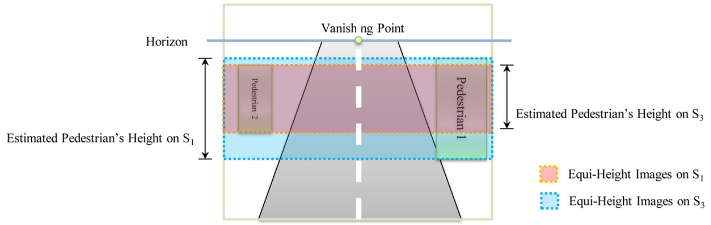

Vanish ng Point
Horizon
Pedestrian 2
Pedestrian 1
Estimated Pedestrian's Height on S_1
Estimated Pedestrian's Height on S_3
Equi-Height Images on S_1
Equi-Height Images on S_3

Fig. 2. Equi-height image generation on a 2D image [1]

generate the SVM classifier using a modified dataset and perform the detection using a trained classifier. After finishing detection, we perform tracking on the equi-height mosaicking image of the next frame.

3.1 Detection

Training for Equi-Height Image. In order to generate the SVM classifier, we make a modified dataset suitable for the equi-height mosaicking image. To generate this modified dataset, we gather various pedestrian data of the size 64x128 and perform reflection. Next, we make the tall pedestrian data by cropping and resizing the normal data. We also make the small pedestrian data using padding and resizing. Because some pedestrian heights on the equi-height image can be taller or smaller than the average height, we generate both tall and small pedestrian datasets. Fig. 3 shows an example of modified datasets created an INRIA person dataset [12].

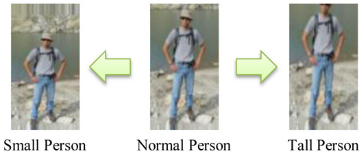


Fig. 3. Modified training dataset as pedestrian's height

Detection on Equi-Height Mosaicking Image. In this step, we use the HOG [2] feature to generate the 1D search based SVM classifier for pedestrian detection. In order to function in real-time, the proposed system applies a modified method based on the existing GPU-based HOG detector [3,13]. It is fast enough to be used in real-time but if we want to detect objects of various scales, the existing HOG detector must become very slow in order to process the image scaling. Our modified method focuses on reducing the computational complexity of the sliding window based HOG detector because it spends too much execution time on image scaling and linear SVM classification [3]. In order to reduce the computational complexity, we use an equi-height mosaicking image that has a fixed image height. This changes the search process of the sliding window to 1D search without any additional scaling issues. Additionally, the proposed system reduces false positive rate by restricting the computational region on the captured image. The magnified image in Fig. 4 shows the detected result using a 1D search on an equi-height mosaicking image. The green rectangles represent the detected regions of the pedestrian.

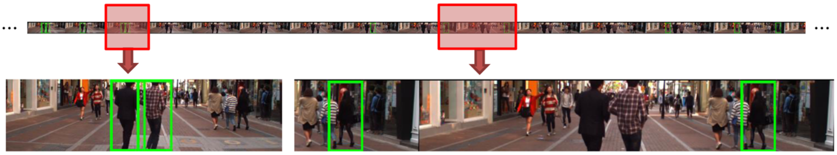


Fig. 4. Detected results on an equi-height mosaicking image

3.2 Tracking and Non-maximal Suppression

After the pedestrian detection process is finished, we track detected pedestrians. To ensure efficient tracking, we perform 1D matching on equi-height mosaicking images between time $t-1$ and time t. In this case, we use the HOG feature computed in the previous step. At this time, we match only neighbor blocks of the detected position on the equi-height mosaicking image of time t frame because pedestrians may move slowly. This approach also reduces the computational complexity of the proposed system. Fig. 5 shows the tracking algorithm performed on an equi-height mosaicking image. The red, green and blue rectangles represent the detected region, tracked region and matching candidate area, respectively.

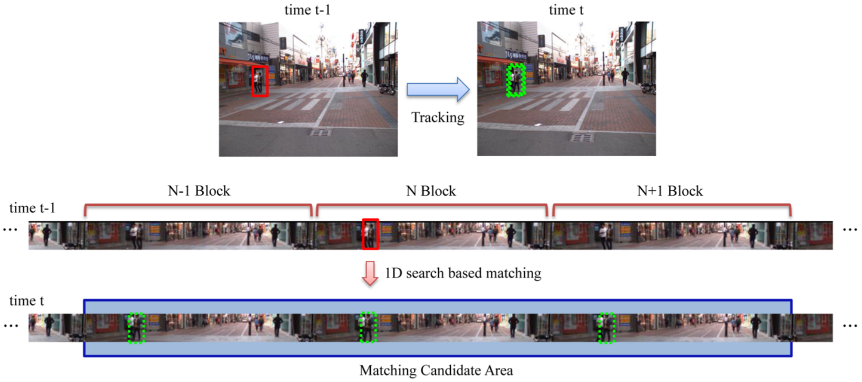
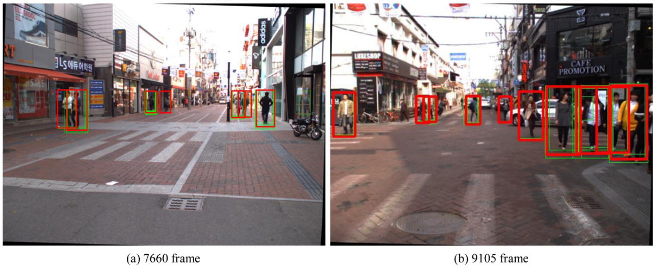


Fig. 5. Tracking result using 1D matching on an equi-height mosaicking image

Finally, we map the detected rectangles onto the original image and perform non-maximal suppression [14] to group the detected multiple overlapping pedestrians. Fig. 6 shows the results of mapping on the original image and the non-maximal suppression. Green rectangles are pedestrian candidates that are converted to the original image while red rectangles are definitive pedestrians after non-maximal suppression.


(a) 7660 frame

(b) 9105 frame

Fig. 6. Final detected results

4 Experiments

4.1 Experiment Setup

Our experiments demonstrate the efficiency of a pedestrian detection system using equi-height mosaicking images, originally designed for vehicle detection [1]. The proposed system executes faster than the existing GPU based HOG+SVM detector under the same experimental environment because it reduces the computational complexity of the detection and tracking processes by using equi-height mosaicking images. In these experiments, we compare the performance of the proposed system with that of a GPU-based OpenCV HOG+SVM detector [13]. In order to measure processing time, we use Visual C++ on an i5 750 CPU with 16GB of RAM and nVidia Geforce GTX 680. The experiment is carried out on the road of a crowded downtown area in day-time. To train the SVM classifier, we have used a training data set that includes positive images of 7,248 pedestrians and negative images of 27,135 non-pedestrians. The accuracy of our detection is measured with an F1 measure [15].

4.2 Experimental Results

As shown in Table 1, our approach enhances the execution time in whole steps. In particular, image resizing and block histogram computation are improved by restricting the detecting area. Precision has improved but only marginally so. Because both systems used a linear SVM classifier with the same performance level, their accuracy is also equal. Therefore, the precision of the proposed system seems to have improved because the equi-height mosaicking image eliminates unnecessary image areas using geometric information of the road scene. However, the proposed system performs faster than the GPU-based OpenCV HOG+SVM detector under the equal sampling level. Therefore, the proposed system can be added to other hypothesized methods of improving detection accuracy.

Table 1. Experimental results

Approach	Size	Sampling Level	Execution Time(ms)			Detection rate(%)		
			Resizing	Block Histo.	Total	Precision	Recall	F1
HoG+SVM	640 x 480	40	16	245	348	77.48	72.26	74.78
Ours			26	25	65	90.47	76.66	83.00

5 Conclusion

In this paper, we described a pedestrian detection and tracking system using equi-height mosaicking images in real-time. Our approach is an improvement on the execution time of the existing GPU-based acceleration approach. In order to improve its execution time, it reduces computational complexity by applying a 1D search for pedestrians on equi-height mosaicking images. Further, it applies an efficient tracking method using 1D matching on equi-height mosaicking images. These techniques improve processing time without decreasing accuracy. However, accuracy is not necessarily improved. Therefore, we will continue researching ways to reduce the false alarm and missing rate in future work.

Acknowledgments. This work was supported by the Industrial Strategic technology development program, Development of Driver-View based in-Vehicle AR Display System Technology Development(10040927), funded by the Ministry of Knowledge Economy (MKE, Korea) and supported by the MSIP(Ministry of Science, ICT & Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-1005) supervised by the NIPA(National IT Industry Promotion Agency.)

References

1. Park, M.W., Jung, S.K.: Real-time vehicle detection using equi-height mosaicking image. In: 2013 International Conference on ACM Reliable and Convergent Systems. ACM (accepted, 2013)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005)
3. Prisacariu, V., Reid, I.: FastHOG - a real-time GPU implementation of HOG. Technical Report 2310/09, Department of Engineering Science, Oxford University (2009)

4. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2903–2910. IEEE (2012)
5. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(11), 1330–1334 (2000)
6. Safonova, I., Leeb, H., Kimb, S., Choib, D.: Intellectual two-sided card copy
7. Burger, W., Burge, M.J.: Digital image processing. Springer (2008)
8. Nieto, M., Salgado, L.: Real-time robust estimation of vanishing points through nonlinear optimization, pp. 772402–772402–14 (2010)
9. Torr, P.H., Zisserman, A.: Mlesac: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78(1), 138–156 (2000)
10. Park, M.W., Jang, K.H., Jung, S.K.: Panoramic vision system to eliminate driver's blind spots using a laser sensor and cameras. International Journal of Intelligent Transportation Systems Research 10(3), 101–114 (2012)
11. Johnson, J.: The average walking stride length (May 2011), `http://www.livestrong.com/article/438170-the-average-walking-stride-length/`
12. Dalal, N.: Inria person dataset, `http://pascal.inrialpes.fr/data/human/`
13. OpenCV Dev Team: Opencv gpu hog detector (July 2013), `http://docs.opencv.org/modules/gpu/doc/object_detection.html/`
14. Devernay, F.: A Non-Maxima Suppression Method for Edge Detection with Sub-Pixel Accuracy. Technical Report RR-2724, INRIA (November 1995)
15. Yang, Y.: An evaluation of statistical approaches to text categorization. Information Retrieval 1(1-2), 69–90 (1999)

Feature Selection for HOG Descriptor Based on Greedy Algorithm

Yonghwa Choi[1], Sungmoon Jeong[2], and Minho Lee[1,*]

[1] School of Electronics Engineering, Kyungpook National University
1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, South Korea
yhchoi@ee.knu.ac.kr, mholee@gmail.com
[2] School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
jeongsm@jaist.ac.jp

Abstract. In order to make an efficient face recognition algorithm with real time processing, we should design good feature extraction and classification methods by considering both low computational costs and high classification performance. Among various feature extraction methods, the histogram of oriented gradient (HOG) feature shows good classification performance to classify human faces. However, high-dimensional features such as HOG feature waste lot of memory and computational time. some parts of HOG features for occluded face regions have negative effects in classifying face images, especially occluded face images. Therefore, we should select variable HOG features not only to reduce the computational costs but also to enhance classification performance. In this paper, we applied the greedy algorithm to effectively select the good features within traditional HOG feature. In order to compare the proposed feature extraction with the conventional HOG feature, we fixed classification method such as compressive sensing technique for selected features. Experimental results show that the proposed feature extraction has better classification performance than the traditional HOG features for face datasets with partial occlusion and/or various illumination conditions.

Keywords: Feature selection, Greedy algorithm, Histogram of Oriented Gradient, Face recognition.

1 Introduction

An embedded face recognition system such as surveillance system tries to classify a face in a limited environment[1]. A face recognition system should take low memory with less computational load. The theory of feature reduction offers basic principles for working with lower-dimensional measurements of high resolution images without significantly compromising recognition performance.

In recent years, pattern recognition tends to apply the feature reduction method because of its high-accuracy performance. At first, most algorithms extract high-dimensional features and reduce dimension of extracted features by various methods.

[*] Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 417–424, 2013.

Feature selection techniques, such as greedy algorithm [2, 3], best first[4], Scatter search[5], can efficiently reduce feature dimension.

Navneet Dalal and his colleagues [6, 7] proposed the histogram of oriented gradient (HOG) descriptor in pedestrian detection. Authors in [8, 9] successfully applied HOG descriptors to the problem of face recognition. HOG feature is generated from many cells that are part of an image. Most cells are usually not used for classification and classification accuracy is primarily determined by only small number of cells. Moreover, some cells drop classification accuracy. Therefore, high-dimensional features such as HOG feature wastes a lot of memory capacity and computational time. In this regard, we propose a new feature extraction method using HOG descriptor for improving accuracy performance. In this method the proposed algorithm to select features is based on greedy algorithm which is one of the simplest algorithms [3].

The rest of this paper is organized as follows: In Section 2, we describe problems of face recognition. Section 3 presents the proposed algorithm feature selection with greedy algorithm from HOG feature. In Section 4 presents performance evaluation of the proposed algorithm as a face recognition system and finally, we draw our conclusions in Section 5.

2 Problem Statement

2.1 Face Recognition and Its Limitation

In face recognition, different illumination conditions, worn accessories such as glasses and facial expressions are considered noise. Fig. 1 shows face images with in conditions: (c) neutral, (d) smiling, (b) wearing sunglasses. Important feature points are different in these images as per noise level. For example, mouth region appears as noise feature to measure the similarity between neutral face and smiling face as shown in Fig. 1 (f). Similarly, eye region becomes a noise feature in measuring the similarity between neutral face and face with sunglass as shown in Fig. 1 (a).

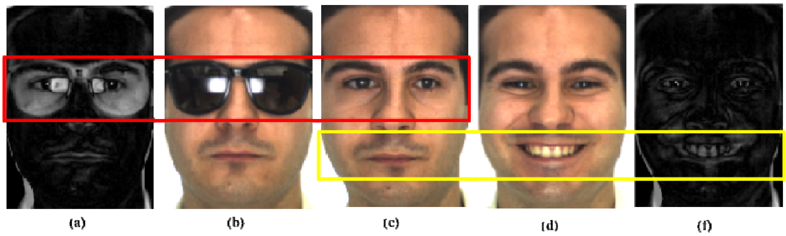


Fig. 1. Examples of noise of facial part: (a) subtraction image between (b) and (c), (b) wearing sunglass face, (c) base face, (d) smiling face, (f) subtraction image between (c) and (d)

Valuable feature points should be different according to given data so that feature selection algorithms can find valuable feature points to extract good feature. Position can be pixel or region on features based on feature selection.

3 Proposed Algorithm

We extract the HOG feature from face images and select valuable features to distinguish each class. For this we use Greedy algorithm, which can easily select most efficient features. Fig. 2 shows a flowchart of the proposed face recognition algorithm. In the training period, we select number of HOG cells by greedy algorithm. Then we can apply the selected HOG feature to classify the test.

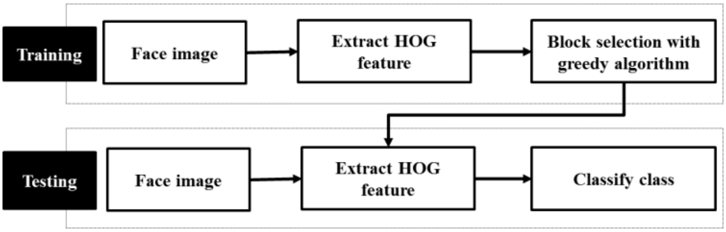


Fig. 2. Flowchart of the proposed algorithm

3.1 Histogram of Oriented Gradient

The HOG is created in minimum units of cell. It is possible to extract good feature of each part by merging them with nearby cells. Most of them are based on grouping cells into larger spatial cells and contrast normalizing each cell, separately. We typically overlap the cells so that each scalar cell's response contributes to several components of the final descriptor vector. HOG descriptor can effectively represent appearance of a face. It is also robust in cases of little noise such as occlusion and shift because of its cell structure. This is one of the most important reasons for us to use HOG descriptor as feature vector.

3.2 Feature Selection with Greedy Algorithm

Greedy forward feature selection (GFFS) algorithm adds one feature dimension at a time to a set of already selected features. The algorithm then checks the goodness of the feature by training and testing classifiers on cross-validation. The best feature, in terms of average accuracy, is then added to the set of selected features. Then the next iteration begins. HOG consists of many cells. To select the best cell, we check classification accuracy of all cells and select the cell with highest classification accuracy. While checking the accuracy in the next iteration, features are extracted by combining candidate cell and selected cell. The threshold is the number of cells giving highest classification accuracy in the training data. Usually, the number of cells is half of the number in all blocks. The following is the summary of our proposed feature selection method based on HOG with greedy algorithm.

Algorithm 1. Greedy algorithm for HOG

```
Input : HOG feature $X$ , number of cell $n$
Output : Selected cell $S$
Steps :
  1) Initialize $S \leftarrow \emptyset$, Generate a position of HOG cell $B$

  2) Calculate classification accuracy
         for all cell at $B$
             $l = arg\ \max_{i \in B} Classifier(X(S \cup B_i))$
         end for

  3) Update parameter $S$, selected cell $B_t$, accuracy performance $C_t$, time $t$
         $S \leftarrow \{S \cup B_l\}$
         $B_t \leftarrow B_l$
         $C_t \leftarrow Classifier(X(S))$
         $t \leftarrow t + 1$

  4) if $B$ is null,
         $S \leftarrow \emptyset$
         $k = arg \max_{0<i \leq n} C_i$
         for $i = 1{:}\,k$
             $S \leftarrow \{S \cup B_i\}$
         end for
     Otherwise repeat steps 2-3
```

4 Experimental Result

The proposed algorithm is compared with conventional HOG algorithm to verify its advantages in terms of classification accuracy and computational time. We performed the experiment in three following conditions: (1) same environment setting for training/test phase (2) partially occluded face images, (3) face images with various illumination conditions.

4.1 Experimental Setup

Aim of the experiment was to show that our proposed greedy feature selection algorithm with HOG feature compressed sensing using QR decomposition [10] can improve classification performance. Our algorithm also reduces the computation time.

We use two face databases. First, AR face database [11] and the Extended Yale B database [12]. The AR database consists of over 4,000 frontal images for 126 subjects. For each subject, 26 pictures were taken in two separate sessions in same condition. These images include facial variations in terms of illumination and expressions. In the experiment, these images were cropped with a dimension of 165 x 120 and converted to gray scale. To extract HOG feature from AR face database, our experiment used 15x15 blocks with an overlap and made cells to merge into four blocks. We extracted the histogram with 8 numbers of bins at 180 degree of angle.

The Extended Yale B database consists of 2,414 frontal-face images of 38 individuals. About 60 pictures were taken in two separate sessions in same condition. These images included only illumination change. In the experiment, images were cropped to 192 x 168 dimensions and the scale was resized to 96 x 84 pixels. For extraction of HOG feature from extended Yale B face database, we used 8x8 blocks with an overlap and made cells by merging into four blocks. Other parameters were same with AR face database. The experiments in all three conditions were performed 10 times. We report the average values as results in table. .

4.2 Block Selection for HOG Feature

To improve classification accuracy, HOG feature is selected from different positions depending on the noise pattern of dataset. Fig. 3 shows the sequence of selection for three-faces: normal, wearing sunglasses and wearing scarf. First 10 cells were represented by red cells, and rest 30 cells were represented by yellow cells. As we discussed earlier, important feature point is different according to noise. Selected cells show most important position to classify.

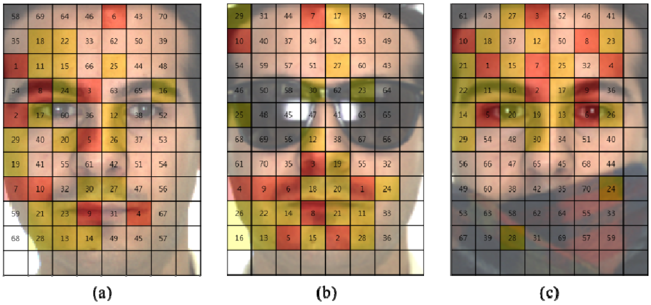


Fig. 3. Selection sequence for three faces: (a) normal, (b) wearing sunglasses, (c) wearing scarf

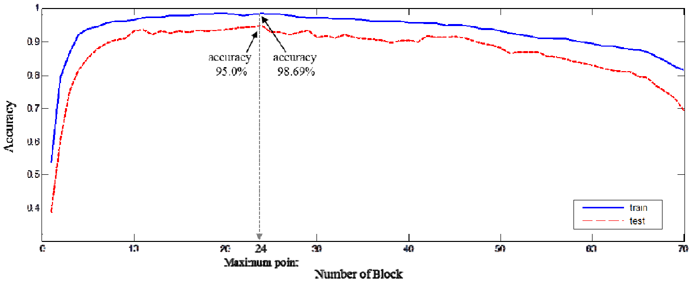


Fig. 4. Decision of number of cells

4.3 Decision on Number of Cells

Number of cells affects classification accuracy. We automatically determine the threshold, which is the number of cells giving highest classification accuracy in the training. Commonly, the number of cells is decided as the half of all blocks. Fig. 4 shows classification accuracy of occluded face images according to growing number of cells in the experiment. . The number of cells having best performance is nearly the same at training and testing. Fig. 4 also shows that all cells do not positively contribute towards increasing the classification accuracy.

4.4 Improvement in Classification Accuracy

Same Environment Setting for Training/Test Phase. In this part of the experiment, we chose a subset of the data set consisting of 50 male subject images and 50 female subjects images and 14 non-occlusion face images in each subject. In the training, we randomly chose 30 classes for selecting cells. Remaining 70 classes were used for testing to check classification accuracy and computational time. Table 1 shows that the proposed feature has better classification performance than the traditional HOG feature. This is despite the fact that our proposed model consists some parts of traditional HOG feature. Moreover, we found that computational time of the proposed model is less than traditional HOG feature.

Table 1. Comparison of accuracy performances and computational time with **same environment setting**: using AR face database

Algorithm	Performance (%)	Number of cell	Time(s)
HOG	98.17±2.12%	70	11.24±1.25
Proposed	99.04±1.05%	42.5±4.2	5.41±0.82

Partially Occluded Face Images. In this part of the experiment, we tested two occluded faces: wearing sunglasses and wearing scarves. In each subject, there were 7 non-occluded face images, 6 wearing sunglasses face images, and 6 wearing scarves face images. Training set consisted of 7 non-occluded faces in each subject. Testing set consisted of 6 occlusion faces in each subject. In the training, we randomly chose 30 classes for selecting cells. Remaining 70 classes were used for testing to check classification accuracy and computational time. Table 2 shows the improved classification accuracy to select the number of cells by the proposed algorithm. Also, the proposed algorithm can greatly reduce computational time by reducing the number of cells.

Table 2. Comparison of accuracy performance and computational time with partially occluded face images using AR face database

Test DB	Algorithm	Performance (%)	Number of cell	Time(s)
Sunglasses	HOG	75.18±7.32%	70	10.13±1.04
	Proposed	95.65±2.50%	22.4±3.1	2.81±0.65
Scarves	HOG	78.81±6.18%	70	11.86±1.21
	Proposed	94.3±3.61%	21.0±2.6	2.19±0.52

Face Images with Various Illumination Conditions. In the training, we randomly chose 10 classes for selecting cells. Remaining 28 classes were used for testing to check classification accuracy and computational time. Table 3 shows that the proposed model has greater classification performance and lesser feature dimensions than the traditional HOG. Even though classification accuracy of the proposed algorithm is same with traditional HOG, the proposed algorithm can still reduce the number of cells significantly.

Table 3. Comparison of accuracy performance and computational time with face images in various illumination conditions using YaleB face database

Algorithm	Performance (%)	Number of cell	Time(s)
HOG	99.82±0.12	99	15.82±1.57
Proposed	99.82±0.12	5±2.2	1.02±0.43

5 Discussion and Conclusion

In this paper, we proposed efficient face recognition system using feature selection with greedy algorithm, and histogram of oriented gradient (HOG). Main idea of the proposed algorithm is to reduce dimension of HOG feature and improve classification performance using only selected HOG descriptor by greedy algorithm. Experimental results showed that proposed algorithm performs better in terms of computational time and accuracy than typical HOG. We have also shown the strength of HOG descriptor using face image in various noise conditions. Through feature selection the HOG descriptor was robust in illumination changing, expression and occlusion.

In our future work, we are considering incremental learning for the recognition of various kinds of faces based on the proposed method.

Acknowledgment. This research was supported by the Converging Research Center Program through the Ministry of Science, ICT and Future Planning, Korea (2013K000333).

References

1. Pun, K.H., Moon, Y.S., Tsang, C.C., Chow, C.T., Chan, S.M.: A face recognition embedded system. In: Defense and Security. International Society for Optics and Photonics (2005)
2. Farahat, A.K., Ghodsi, A., Kamel, M.S.: An efficient greedy method for unsupervised feature selection. In: 11th IEEE International Conference on Data Mining (ICDM). IEEE (2011)
3. Vafaie, H., Imam, I.F.: Feature selection methods: genetic algorithms vs. greedy-like search. In: Proceedings of International Conference on Fuzzy and Intelligent Control Systems (1994)

4. Xu, L., Yan, P., Chang, T.: Best first strategy for feature selection. In: 9th International Conference on Pattern Recognition. IEEE (1988)
5. Glover, F., Laguna, M., Martí, R.: Fundamentals of scatter search and path relinking. Control and Cybernetics 39(3), 653–684 (2000)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE (2005)
7. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
8. Déniz, O., Bueno, G., Salido, J., De la Torre, F.: Face recognition using histograms of oriented gradients. Pattern Recognit. Lett. 32(12), 1598–1603 (2011)
9. Albiol, A., Monzo, D., Martin, A., Sastre, J., Albiol, A.: Face recognition using HOG–EBGM. Pattern Recognit. Lett. 29(10), 1537–1543 (2008)
10. Shi, Q., Eriksson, A., van den Hengel, A., Shen, C.: Is face recognition really a compressive sensing problem? In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2011)
11. Martinez, A.M.: The AR face database. CVC Technical Report, vol. 24 (1998)
12. Lee, K.-C., Ho, J., Kriegman, D.J.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Pattern Anal. Mach. Intell. 27(5), 684–698 (2005)

A Fault Diagnosis Method under Varying Rotate Speed Conditions Based on Auxiliary Particle Filter

Hongxia Pan and Jumei Yuan

School of Mechanical Engineering and Automation,
North University of China, Xueyuan Road No.3, 030051 Taiyuan, China
panhx1015@163.com

Abstract. Varying rotate speed can cause changes in a measured gearbox vibration signal. There is a need to develop a technique to provide accurate state indicator of gearbox under fluctuating rotate speed conditions. This paper presents an approach for gearbox fault detection under varying rotate speed condition based on auxiliary particle filter. Firstly, the model of vibration part which sensitive to the alternating rotate speed condition was established based on the relation of cosine signal three points sampling values. Then this part vibration signal was estimated based on auxiliary particle filter. Based on these the residual signal was obtained which lower sensitiveness to the alternating rotate rate condition. Thus the gearbox fault was detected by the residual signal statistic quantity kurtosis and amplitude of Fourier transform. Finally, the different work condition vibration signals of the laboratory gearbox under varying rotate speed condition were detected and signal processing was studied with those signals as examples. The results show that the proposed method is feasible and effective.

Keywords: Auxiliary particle filter, Gearbox, Fault diagnosis, Varying rotate speed, Vibration signal.

1 Introduction

Gearbox is the equipments of gear transmission for various machineries. It has very important practical significance to diagnosis its fault to ensure its normal operation. Vibration measurement is one of the most common fault diagnosis methods. But conventional techniques for fault detection are based on the assumption that changes in vibration signal are only caused by deterioration of the gearbox. Varying rotate speed can cause changes in a measured gearbox vibration signal. There is a need to develop a technique to provide accurate state indicator of gearbox under fluctuating rotate speed conditions. For this problem, order tracking technique has become one of the important approaches for fault diagnosis in rotating machinery [1~5]. Order tracking technique normally exploits a vibration or a noise signal supplemented with the information of shaft speed for fault diagnosis of rotating machinery. The ordered amplitude figure of analysis gives the information of harmonic order signal in the rotating machinery. Interest in fault diagnosis using order tracking technique has

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 425–432, 2013.

grown significantly with the advance in digital signal-processing algorithm and technology in the last two decades. In general, an order spectrum gives the amplitude of signal as a function of harmonic order and shaft speed. Order tracking is also used to analyze and track the energy of order signal from dynamic signal. However, order tracking may produce frequency smearing in high speed and high order signal. When there is either high sweep rate or low sweep rate, the analysis about order signal is not enough, and this will result in erroneous result in order tracking. Therefore this paper proposed a new fault detection method for gearbox under varying rotate rate condition. That is using auxiliary particle filter to estimate the vibration signal which related to the change of rotate speed. The errors of measure value with the estimated signal as residual signal, which was used to diagnose the gearbox state.

2 State Equation Establishment and State Estimation

2.1 State Equation Establishment

The conventional algorithms used in fault diagnostic techniques fall into two categories. One is Fourier transform with a fixed sampling rate for obtaining frequency domain information; the other is tracking with various sampling rates. The second method employs a re-sampling scheme synchronous with the shaft revolution. The time domain data are hence converted to revolution-domain data. Then the FFT is also applied to obtain the order spectrum with respect to engine speed. Both the time and the frequency resolution of this approach are essentially varied with the shaft speed. This FFT order-tracking method relies on accurate measurement of the tachometer signal. In general, the vibration signal generated by rotating machinery essentially consists of a combination of the basic frequency with narrowband frequency components and its harmonic frequencies, most of which are related to the revolution of the machine. The vibration signal $y(t)$ containing k orders generated by one rotating shaft can be written as[6]:

$$y(t) = A_1 \cos[\theta(t) + \phi_1] + \cdots + A_k \cos[k\theta(t) + \phi_k] \tag{1}$$

$$\theta(t) = \int_0^t \omega(\tau) d\tau = \int_0^t 2\pi f(\tau) d\tau \tag{2}$$

Where: A_k is amplitude of kth order, ϕ_k is phase of kth order, $\theta(t)$ is angular displacement of rotating gear computed by the following integral, $\omega(\tau)$ is instantaneous angular frequency of the rotating shaft.

Any of constant frequent cosine signal at three points sampling values to meet the equation:

$$x(n\Delta t) - 2\cos(\omega\Delta t)x((n-1)\Delta t) + x((n-2)\Delta t) = 0 \tag{3}$$

Where, $x(n\Delta t)$ is the n moment sampling value; Δt is the interval of sampling; and ω is the instantaneous angular frequency of cosine signal.

When gearbox rotate rate occur change the instantaneous angular frequency ω wound varying. Equation (3) wound change to follow:

$$x(n\Delta t) - 2\cos(\omega\Delta t)x((n-1)\Delta t) + x((n-2)\Delta t) = \xi \tag{4}$$

Where ξ is the change in the relationship between cosine signal three points value caused by the change of instantaneous angular frequency ω.

Selecting state variables [$x(n\Delta t)$ $x((n-1)\Delta t)$ $]^{\mathrm{T}}$, then equation (4) can be expressed as:

$$\begin{bmatrix} x(n\Delta t) \\ x((n-1)\Delta t) \end{bmatrix} = \begin{bmatrix} 2\cos(\omega\Delta t) & -1 \\ 1 \quad 0 & \end{bmatrix} \begin{bmatrix} x((n-1)\Delta t) \\ x((n-2)\Delta t) \end{bmatrix} + \begin{bmatrix} \xi \\ 0 \end{bmatrix} \tag{5}$$

The corresponding measurement equation is expressed as:

$$y(n) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x(n\Delta t) \\ x((n-1)\Delta t) \end{bmatrix} \tag{6}$$

Assumption that the number of periodic components we model is q among the gearbox vibration signal. the Those high order harmony periodic components together with model error, measurement error and noise are regarded as system noise. Then the state equation of gearbox vibration signal can be established as:

$$\begin{bmatrix} x_1(n) & x_1(n-1) & x_2(n) & x_2(n-1) & \cdots & x_q(n) & x_q(n-1) \end{bmatrix}^T =$$

$$\begin{bmatrix} 2\cos(\omega_1\Delta t) & -1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 2\cos(\omega_2\Delta t) & -1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 2\cos(\omega_q\Delta t) & -1 \\ 0 & 0 & \cdots & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(n-1) \\ x_1(n-2) \\ x_2(n-1) \\ x_2(n-2) \\ \vdots \\ x_q(n-1) \\ x_q(n-2) \end{bmatrix} \tag{7}$$

$$+w(n)$$

The corresponding measurement equation is:

$$y(n) = \begin{bmatrix} 1 & 0 & 1 & 0 & \mathrm{L} & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(n) & x_1(n-1) & x_2(n) & x_2(n-1) & \mathrm{L} & x_q(n) & x_q(n-1) \end{bmatrix}^T + v(n) \tag{8}$$

Where: $w(n)$ is $2q$ dimensions state noise, $v(n)$ is sum of ξ in equation (4) and high order harmony periodic components together with measurement error, noise and so on.

2.2 State Estimation Based on Auxiliary Particle Filter

Particle filtering is a method for state estimation that is not dependent on the probability density function (pdf) of the measurements. In the general case the equations of the optimal filter used for the calculation of the state-vector of a dynamical system do not have an explicit solution. This happens for instance when the process noise and the noise of the output measurement do not follow a Gaussian distribution. In that case approximation through Monte-Carlo methods can be used. A sampling of size N is assumed, i.e. N i.i.d. (independent identically distributed) variables ξ1, ξ2, … ,ξN. This sampling follows the pdf $p(x)$. i.e. ξ1: N ~ $p(x)$. Instead of $p(x)$ the function $p(x) \approx p^N(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi i}(x)$ can be used. It is assumed that all points ξi have an equal weighted contribution to the approximation of $p(x)$. A more general approach would be if weight factors were assigned to the points ξi , which would also satisfy the normality condition $\sum_{i=1}^{N} w^i = 1$. In the latter case

$$p(x) \approx p^N(x) = \sum_{i=1}^{N} w^i \delta_{\xi i}(x) \tag{9}$$

If $p(\xi^i)$ is known then the probability $p(x)$ can be approximated using the discrete values of the pdf $p(\xi^i) = \omega^i$. If sampling over the pdf. $p(x)$ is unavailable, then one can use a pdf $q(x)$ with a similar support set, i.e. $p(x) = 0 \Rightarrow q(x) = 0$. Then it holds that $E(\phi(x)) = \int \phi(x)p(x)dx = \int \phi(x)q(x)\frac{p(x)}{q(x)}dx$. If the N samples of $q(x)$ are available at the points $\xi^1,\xi^2,\ldots,\xi^N$, and the weight coefficients ω^i are defined as $\omega^i = \frac{p(\xi^i)}{q(\xi^i)}$, then it is easily shown that

$$E(\phi(x)) \approx \sum_{i=1}^{N} \omega^i \phi(\xi^i) \tag{10}$$

where: $\xi^{1:N} \sim q(x)$, $\omega^i = \frac{p(x^i)}{q(x^i)}$.

Eq. (10) is important: assume that the pdf $p(x)$ is unknown (target distribution), however the pdf $q(x)$ (Importance law) is available. Then, it is sufficient to sample on $q(x)$ and find the associated weight coefficients ω^i so as to calculate $E(\phi(x))$ [7].

To implement particle filters, a number of issues need to be considered, including degeneracy, the selection of the importance density, and the number of particles required. These issues are now discussed briefly. Degeneracy is where, after a number of time points, only one particle has significant weight. Thus, considerable computational effort is expended on updating particles whose contribution to the approximation of $p(x)$ is negligible. Re-sampling has been used as a standard procedure to resolve this problem. Through re-sampling the weights are re-set to 1/N as the particles are independent and identically distributed and drawn from a discrete density function. By re-sampling, the particles with small weights will be eliminated. The second issue is how to select the importance density. One approach is to use prior distribution, which will yield a simple form for updating the weights. However, this importance density may be sensitive to the presence of outliers, and can be improved if it depends on the current measurement. This idea was further developed by Pitt and Shephard who proposed that u_k^i, the mean of $p(x_k \mid x_{k-1} i)$, is first calculated and then the importance density is redefined as:

$$q(x_k \mid z_k) \propto p(z_k \mid u_k^i) p(x_k \mid x_{k-1}^i) \omega_{k-1}^i \tag{11}$$

By utilizing u_k^i, new particles are generated from particles at the previous time point, conditional on the current measurement z_k, which will be closer to the true states. The calculation of weights is thus given by:

$$\omega_k^i \propto \frac{p(z_k \mid x_k^i)}{p(z_k \mid u_k^i)} \tag{12}$$

Particle filters with this importance density and re-sampling step are termed Auxiliary Sampling Importance Re-sampling (ASIR) filters[8]. ASIR filters are used in this study to estimate the gearbox vibration signal which sensitive to the change of rotate speed. The residual signal was obtained by the measure data minus estimated value. Then the gearbox fault was detected by the residual signal statistic quantity kurtosis and amplitude of Fourier transform.

3 Experimental Set-Up

Automaton The vibration data used in this paper were obtained from our laboratory gearbox experimental device. It has three axes. Its inner structure shows in Fig.1.

Where z1=30,z2=69, z3=18, z4=81. bearings of input shaft and middle shaft are 6406E, and bearings of output shaft are 6312E. This experimental used of Yangzhou radio two Plant production CA-YD piezoelectric accelerometers. Accelerometers were installed through adhesive, and use a screw to fix the acceleration sensor. Here choose six measuring points to carry on vibration acceleration signal gathering. The first measuring point located in the right bearing department of the box input shaft, measured the axis of vertical plane of a vibration. Sensor model is YD-81D piezoelectric acceleration sensor, and its sensitivity is 52.24pc/g. The second measuring point located in the right bearing department of the middle box axis, measured the axis of the vertical plane of a vibration. Sensor model is YD-81D piezoelectric accelerometer, and its sensitivity is 49.4pc/g. The third measuring point located in the right bearing department of the box output shaft, measured the axis of the vertical plane of a vibration. Sensor model is YD-81D piezoelectric accelerometer, and its sensitivity is 46.26pc/g. Fourth measurement points is in the left side after the gear (i.e gear with the second axis), measured horizontal plane of vibration which vertical the middle axis. Sensor model is YD-12-45 piezoelectric acceleration sensor, and its sensitivity is 54.91pc/g. Fifth measurement points located in the left side of second shaft bearing, measured horizontal plane of vibration which vertical the second axis. The sensor model is YD-81D piezoelectric accelerometer, and its sensitivity is45.97pc/g. 6th measurement points is in the left anterior side of the cover in box office two shaft bearings, measuring the second axis perpendicular to the vertical plane of vibration. Sensor model is YD-81D piezoelectric accelerometer, and its sensitivity is 50.48pc/g. In addition, the speed pulse, torque pulse, speed signal and torque signal are measured.

The vibration acceleration signals were collected under normal; intermediate haft bearing outer ring falling, intermediate haft bearing inner ring falling and retainer break conditions separately. Adjust the speed of the input shaft by hand so it changes from small to large and from large to small repeatedly. Sampling frequency is taken as 20kHz. Record these values of the acceleration signal. Vibration acceleration units are m/ss.

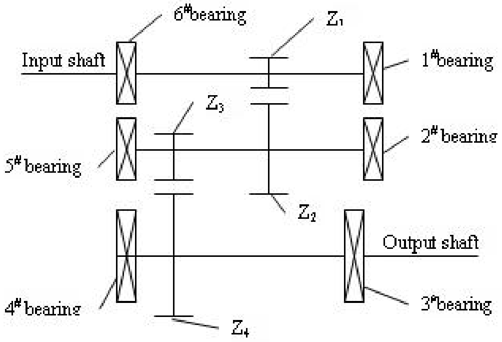
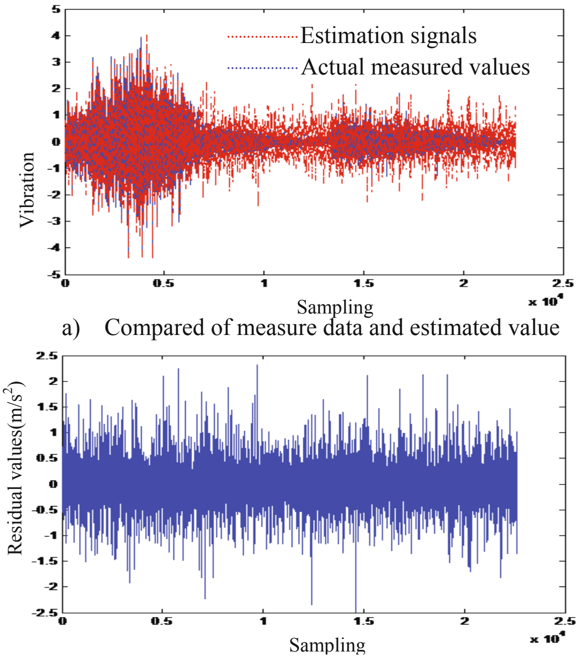


Fig. 1. Gearbox Inner Structure

4 Application Example

Here, only the first measuring point data were analyzed. Using from 1 to 4 harmonics of input shaft, intermediate shaft and output shaft as periodic component. Using equation (7) and (8) as the equations of state vibration signals and measurement equation respectively. The number of particles selected 100. Estimating these signals based on auxiliary variable particle filter. Estimated results of typical working state for retainer break show in Fig.2.



a) Compared of measure data and estimated value

b) The residual signal for retainer break

Fig. 2. Results of estimated for retainer break state

In Fig.2, the charts a) expressions the compared of measure data and estimated value, where the blue represents the actual measured values, the red for the results of estimation signals. The charts b) expressions the residual signal for every state. We can see that the influence of rotational speed change to vibrate signal can be responded by estimated signal goodly in the signal detection range, the residual signal is stationary signals, and have not relationship with the change of rotate speed. It illustrates that the residual signal can be used as a gearbox fault detection signal.

Select the feature value of normal condition residual signal variance, kurtosis and spectrum amplitude as criterion value. Based on this criterion value we can calculate the other condition related feature values. These results show in table 1.

Table 1. Various state relative feature value

State	normal	retainer	bearing	bearing
feature		break	inner fault	outer fault
variance	1	1.0921	1.0641	1.0734
kurtosis	1	1.1662	1.1844	1.2784
amplitude	1	1.0903	1.2089	1.1643

We can see that these relative feature values are all greater than 1. It indicates that compared with the normal condition, the residual signal fluctuation is relatively large under gearbox failure condition. So we can detect gearbox fault online according to the relative characteristics value.

5 Conclusion

Gearboxes often operate under fluctuating rotate speed conditions during service. The fault diagnosis technique becomes more complicated when gearbox is subject to varying operating condition. This paper proposes an approach for gearbox fault detection under varying rotate speed condition based on auxiliary particle filter. The model of vibration part which sensitive to the alternating rotate speed condition was established based on analysis of gearbox vibration signals. This part vibration signal was estimated based on auxiliary particle filter. On this foundation, the residual signal was obtained which is stationary and insensitive to the varying rotate speed. Their features values relate to normal condition were used to detect the gearbox state. Finally, an example analysis shows that the method is feasible and effective. Further work is to further analysis the residual signal and to construct a more effective statistic for fault diagnosis.

Acknowledgment. The work described in this paper has been supported by National Natural Science Foundation of China under the grant No. 51175480 and Natural Science Foundation of Shanxi under the grant No. 2011011019-1. The authors would like to express their gratitude for the support of this study.

References

1. Lin, J., Qu, L.: Feature Extraction Based on Morlet Wavelet and Its Application for Mechanical Fault Diagnosis. J. Sound Vib. 234, 135–148 (2000)
2. Biswas, M., Pandey, A.K., Bluni, S.A., Samman, M.M.: Modified Chain-code Computer Vision Techniques for Interrogation of Vibration Signatures for Structural Fault Detection. J. Sound Vib. 175, 89–104 (1994)
3. Shibata, K., Takahashi, A., Shirai, T.: Fault Diagnosis of Rotating Machinery through Visualization of Sound Signals. Mech. Syst. Signal Proc. 14, 229–241 (2000)
4. Chen, Y.D., Du, R., Qu, L.S.: Fault Features of Large Rotating Machinery and Diagnosis Using Sensor Fusion. J. Sound Vib. 188, 227–242 (1995)
5. Gelle, G., Colas, M., Serviere, C.: Blind Source Separation: A Tool for Rotating Machine Monitoring by Vibration Analysis. J. Sound Vib. 248, 865–885 (2001)
6. Wu, J.D., Bai, M.R., Su, F.C.: An Expert System for the Diagnosis of Faults in Rotating Machinery Using Adaptive Order-tracking Algorithm. Expert Systems with Applications 36, 5424–5431 (2009)
7. Gerasimos, G.R.: Particle and Kalman Filtering for State Estimation and Control of DC Motors. ISA Transactions 48, 62–72 (2009)
8. Pitt, M.K., Phard, N.S.: Filtering via Simulation: Auxiliary Particle Filters. Journal of the Americal Statistical Association 94(2), 590–599 (1999)

Efficient Traffic Sign Detection Using Bag of Visual Words and Multi-scales SIFT

Khanh-Duy Nguyen[1], Duy-Dinh Le[2], and Duc Anh Duong[1]

[1] Multimedia Communications Lab, University of Information Technology,
VNU-HCM, Ho Chi Minh City, Vietnam
{khanhnd,ducda}@uit.edu.vn
[2] National Institute of Informatics, Tokyo, Japan
ledduy@nii.ac.jp

Abstract. Automatic traffic sign detection is important in many applications such as GPS based navigation systems, advanced driver assistance systems, and self-driving cars. Recently, several researches have shown that bag of visual words (BoVW) method is really an interesting and potential choice for this detection problem. However, it is difficult for using this approach in practice due to the high computational cost. To find the exact boundaries of objects, this approach has to scan a large number of image sub-windows over location and scale (e.g. there are approximately 60,000 32x32 pixels sub-windows for an 320x240 pixels image). In this paper, we propose an efficient approach, which use multi-scales SIFT features and coarse-to-fine search strategy, to improve speed of BoVW. We argue that multi-scales SIFT features can be used for quickly detecting the coarse boundaries of objects. Then, the further searching stage only need to concentrate on these discovered boundaries. By this way, the number of image sub-windows is efficiently reduced. The experimental results show that our proposed method significantly improves detection speed without trading off performance.

Keywords: Traffic sign detection, Bag of visual words, Multi-scales SIFT, Sub-windows search.

1 Introduction

The traffic sign detection is a critical task in several applications and research projects, such as: GPS based navigation systems, advanced driver assistance systems, and self-driving cars. Although several studies have been done recent years [1–3], many challenges remain such as: occlusion caused by trees or other objects; cluttered background; bad weather conditions; damaged or faded traffic signs; camera motion blur. These studies always use color as a clue to narrow down the search space. After that global features (i.e. edge images) and shape detectors are used to detect traffic signs. Unfortunately, this approach is not effective in practical conditions.

Besides, recently several researchers have successfully employed a new object detection and recognition approach, Bag of Visual Words (BoVW) [4], which

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 433–441, 2013.

is a state-of-the-art method in image classification [5, 6]. The main idea of this approach is using part-based representation and machine learning algorithms to detect objects. Outstanding contribution of this approach is its effectiveness for occlusion, cluttered background, and damaged objects.

Consequently, BoVW is an interesting and potential choice for solving the traffic sign detection problem. However, this has not been done yet. There are few approaches using BoVW for detection problem in general. The reason here is its low speed. The only one appropriate localization method for BoVW model, sub-windows search, requires a lot of computation. Although several works have been conducted to speed up sub-windows search (e.g. ESS [7]), they are not enough feasible in some cases (see more details in Sect. 2).

In this paper, we proposed an efficient approach to improve speed of sub-windows search. The original idea is inspired from the robustness of invariant-scales features in detection. We argue that multi-scales SIFT [8, 9] can be used for quickly detecting a coarse boundary of objects, after that the further searching stage only need to concentrate on this discovered boundary. To evaluate this approach, we conduct some experiments on the Germany Traffic Sign Detection Benchmark [10]. These experiments are designed to clarify:

- How robust are scale-invariant descriptors SIFT [11] in object detection when object scales significantly change?
- How much cost can be reduced by using multi-scales SIFT instead of SIFT?
- The performance of proposed method comparing to traditional sub-windows search and BoVW.

The results show that our method significantly improves detection speed without trading off performance, even outperforms in some cases.

In the following, several major approaches in speeding up object detection are briefly reviewed and our approach is introduced. Then, we describe our coarse-to-fine search method using multi-scales SIFT in Sect. 3. In Sect. 4, experiments are presented and the results are discussed. Finally, conclusions are drawn in Sect. 5.

2 Previous Works

Recently, there are a lot of efforts to improve detection speed without trading off quality, as shown in [12]. In this section, we only discuss about several approaches that directly relate to our proposal in this paper.

Firstly, one of the most used approaches in object detection has been proposed by Viola & Jones. Their framework use cascades of classifiers to detect object quickly. Unfortunately, this scheme not suitable for BoVW. The main challenge is that it need of a lot of very fast computed features (i.e. Haar-like features), while computational cost of BoVW is very high. However, the idea of cascade of classifier inspires our approach in this paper. The computation cost will be significantly reduced if we can efficiently remove background from the search space.

Another well-known algorithm is efficient sub-windows search - ESS [7]. This algorithm uses brand and bound strategy for finding location of objects quickly. To do that, ESS requires a bound function which has to be developed separately for different classifier kernels. To increase the speed, regions which have maximum bound values will be priority. Unfortunately, the bound function is not simply to develop, especially in case of multi-kernels classifiers. Moreover, some other challenges can trade off the quality as followings:

- Because of misclassifying between positive and negative words, bound function may be inaccuracy. So size and location of objects may not be detected exactly.
- There is only one region which is detected. Therefore, it cannot take advantages of having many overlapped windows, which are usually used to reduce false positive detection.

Third approach is using Histogram integral [13] and Classifier integral [14] to improve computation speed. When histogram integral allows computing histogram of visual words effectively, classifier integral is based on approximate calculations and only works with linear kernel.

Our approach in this paper is inspired from the different idea of coarse-to-fine search using invariant-scale features. If the invariant-scale descriptors are is really robust with scale space, we can find coarse boundary of objects effectively by using sub-windows search with large-size windows. After that, further searches are conducted on this discovered boundary to find exactly positions and sizes of objects. This significantly reduces computation cost comparing to search in the overall image with many sizes of shift windows.

To implementation this idea, we use SIFT, the state-of-the-art scale-invariant features. To be scale-invariant, SIFT is usually used with several scale invariant detectors (Harris-Laplace, Hessian-Laplace, DoG) [6, 11]. However, due to only a few keypoints could be satisfy these detectors, especially in case of low contrast, the classification lack of geometry information, and is only based on several keypoints. Consequently, the performance may decrease. Moreover, in case of detection, size of object may be small (for example: 16 x 16 pixels traffic signs in the GTSDB benchmark). Then, the detection will be failed if there are very few keypoints detected. To overcome this, several works extract SIFT on dense grid, but the drawback is that then SIFT will lose “scale invariant” property. Therefore, instead of single-scale SIFT, multi-scales SIFT is extracted [8, 9].

3 Detection Using Multi-scales SIFT

3.1 Extract Multi-scales SIFT

In the early work [9], multi-scales SIFT descriptors are computed on a regular grid with spacing M pixels. At each grid point the descriptors are computed over several circular support patches with different radii. However, these descriptors are treated independently of each other. Pseudo codes of this algorithm are shown as following:

```
**Input**: Image: $I$, Scales list: $S$
**Output**: multi-scales SIFT descriptors: $Frames$, $Descrs$
Loop $S_i$ in S
    $I'$ = smooth($I$, $S_i$);
    $Frames(i)$, $Descrs(i)$ = extractDenseSIFT($I'$);
End Loop
```

Algorithm 1. Extract multi-scales SIFT descriptor on a regular grid [9]

A more sophisticated approach is proposed in [8]. This work assume that SIFT descriptors computed at multiple scales of the same point span a linear subspace. Then a subspace is build for representing multi-scales SIFT descriptors extracted from each point. Consequently, the distance between a pair of pixels can be measured by the distance between the corresponding subspaces. A new scale-invariant descriptor named Scale-Less SIFT (SLS) is also produced by applying a subspace-to-point mapping algorithm.

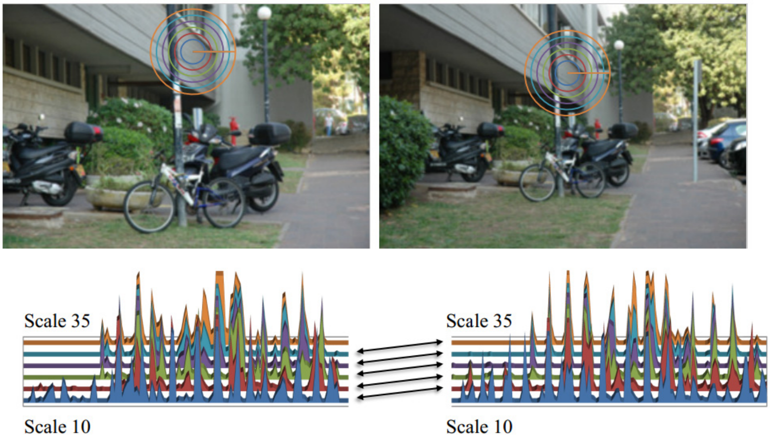


Fig. 1. SIFT descriptors are extracted at a low contrast area where no interest point was detected, at scales ranging from 10 to 35 [8]

3.2 A Coarse-to-Fine Sub-Windows Search Method Using Multi-scales SIFT

Our proposed method consist 2 stages:

- Stage 1 - “coarse finding”: sub-windows search algorithm is used on a coarse grid with multi-scales SIFT descriptors. Due to the robustness of multi-scales SIFT; background regions can be removed efficiently from search space while coarse boundaries of objects remain. Because classifying scores of positive coarse region may be as not good as exactly object region, SVM classifier

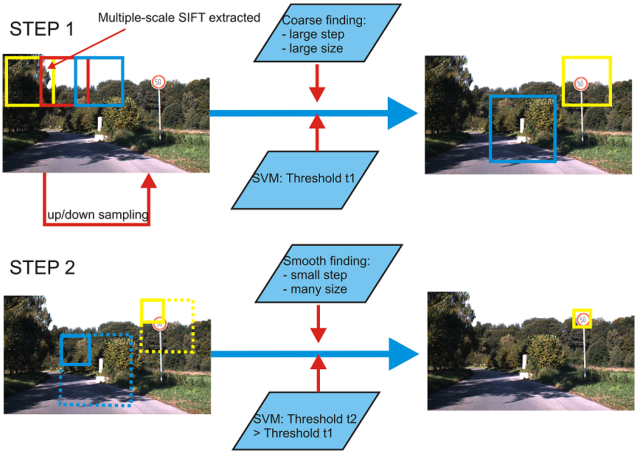


Fig. 2. Overviews of proposed coarse-to-fine sub-windows search method based on multi-scales SIFT descriptors

threshold for positive samples is set at a low value to ensure that objects could not be missed. In our implementation, this value is the default value of SVM (0).

- Stage 2 - "smooth finding": sub-windows search algorithm is used on a smooth grid to find exactly the locations. For this stage, we only find the objects in positive coarse region from step 1. In each coarse region, a randomize process are continuously executed to change steps and sizes of shift windows. Finally the best score windows will be selected. On the other hand, SVM classifier threshold for positive samples is increased to improve the detection accuracy. In our implementation, this value is considered as a parameter of the algorithm.

4 Experiments

We evaluated our method on the GTSDB [10], a standard benchmark for traffic sign detection. According to this benchmark, performance is measured in term of area-under-curve (AUC). For comparing to our proposed method, the performance of the traditional sub-windows search on this benchmark is also evaluated.

4.1 GTSDB Benchmark

The competition task is a detection problem in natural traffic scenes. Participating algorithms need to pinpoint the location of given categories of traffic signs (prohibitory, mandatory or danger).

The performance is computed by an area-under-curve measure for the detector's precision-recall plot on the test dataset. True positives are defined using the following measure:

$$jaccCoeff = \frac{areaIntersection}{groundtruthRoiSize + detectedRoiSize - areaIntersection} \tag{1}$$

and this $jaccCoeff$ has to greater than 0.6.

The benchmark comprises 600 training images and 300 testing images (1360 x 800 pixels) in PPM format. The images contain zero to three traffic signs. The sizes of the traffic signs in the images are vary from 16 x 16 pixels to 128 x 128 pixels. Traffic signs may appear in every perspective and under every lighting condition.

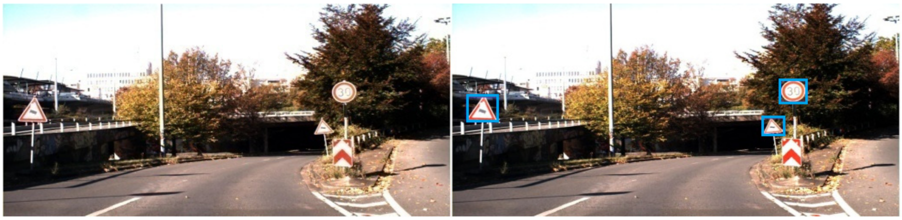


Fig. 3. Input (left)/output (right) sample for the GTSDB benchmark

4.2 Original Sub-windows Search

In this experiment, intensity SIFT (128 bins) descriptors are computed on a dense grid with spacing 6 pixels. These descriptors are clustered using kmeans++ to create a codebook of 1000 words.

In the training SVM model stage, traffic signs images are normalized to the size of 32 x 32 pixels before processed. We also use a bootstrapping process to improve the learning SVM classifier model.

To find the locations of traffic signs, a 32 x 32 pixels shift window is used to scan overall images with spacing 2 pixel. The region in this window is represented as a histogram of 1000 words, then RBF-SVM classifier is used to classify whether traffic signs are present or not. Traffic signs may have many sizes, so we rescale images to different size with several scale factors: *1.5, 1.5/1.25, 1.5/1.25*^{2}*,..., 1/1.25*^{8}.

The results are shown in Table 2 (denseSIFT). For all three category of GTSDB, method using BoVW and subwindows search outperforms the other based-line methods mentioned by this benchmark, as shown in Table 1. This results show the effectiveness of BoVW for traffic sign detection problem. Howerver, as we said in section 1, due to the weakness of detection speed, BoVW has not been an attractive approach so far.

Table 1. Detection accuracy of several base-line methods on the GTSDB benchmark

BASE-LINE METHODS (INI-RTCV)	**AREA UNDER CURVE (%)**	**AREA OVERLAP (%)**
Hough-like Voting Scheme (Prohibitive)	26.09	76.23
Viola-Jones (Prohibitive)	**90.81**	**87.85**
HOG + LDA (Prohibitive)	70.33	78.13
Hough-like Voting Scheme (Danger)	30.41	68.06
Viola-Jones (Danger)	**46.26**	**84.48**
HOG + LDA (Danger)	35.94	78.94
Hough-like Voting Scheme (Mandatory)	12.86	78.46
Viola-Jones (Mandatory)	**44.87**	**88.22**
HOG + LDA (Mandatory)	12.01	77.05

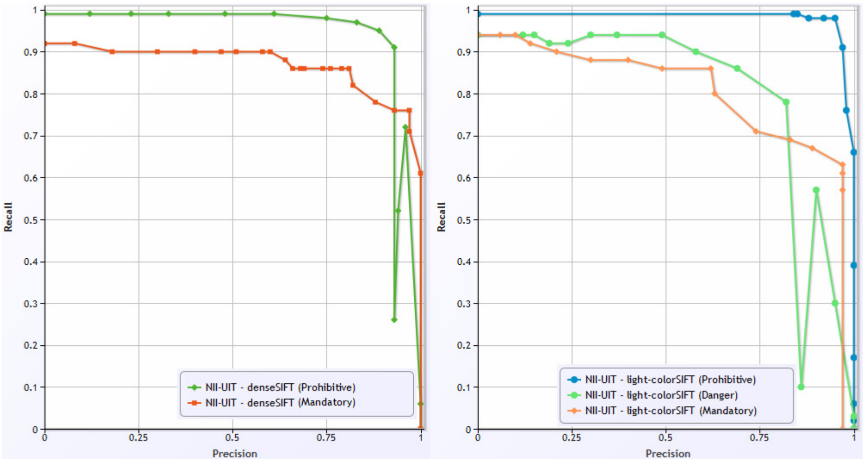


Fig. 4. Precision/recall curves of traditional sub-windows search method (left) and our coarse-to-fine search method (right) in several traffic sign categories

Table 2. Detection accuracy of our coarse-to-fine search method comparing with traditional sub-windows search method

METHOD	**AREA UNDER CURVE (%)**	**AREA OVERLAP (%)**	**Detection time (sec/image)**
denseSIFT (Prohibitive)	94.65	87.48	719
denseSIFT (Mandatory)	**86.97**	**81.77**	
Light-colorSIFT (Prohibitive)	**98.11**	**81.71**	309
Light-colorSIFT (Mandatory)	80.03	83.65	
Light-colorSIFT (Danger)	**80.58**	**79.34**	

4.3 Sub-windows Search Using Multi-scales SIFT

Instead of single-scale SIFT, multi-scales color SIFT descriptors (HSV) are extracted on dense grid with spacing 4 pixels. Similar to the experiment in section 4.1, a codebook of 1000 words is created from these descriptors.

The coarse subwindows search is executed with the same window size. Howerver, the spacing between each window is 8 pixels instead of 2 pixels in section 4.1. This means that the search space will approximately descrease 64 times. Moreover, we rescale image with coarser scale factors: *1.5, 1.5/1.5, 1.5/1.5*^{2}, *...,1.5/1.5*^{5}.

In smooth search step, we use several random window sizes and steps to find exaclty location and size of traffic signs.

Table 2 and Fig4 compares our coarse-to-fine subwindows search (LightcolorSIFT) with the traditional subwindows search (denseSIFT). Our proposed method outperform for prohibitive category (98.11% vs 94.65%), but its performace is lower for mandatory category (80.03% vs 86.97%). The different results between the two category is due to the quality of computed multi-scales SIFT. For the mandatory traffic signs, the sign color is almost solid and the contrast is low. Therefore, although SIFT descriptors are computed on several scales, they are rather similar. Consequently the coarse search step is not effective. In term of detection speed, our coarse-to-find search always shows significantly improvements. Although we use color SIFT (384 bins) instead of intensity SIFT (128 bins) and 1x1+2x2 spatial grid, which increase much computational cost, the speed of coarse-to-find search are 2 times faster than traditional method.

5 Conclusion

We have presented a method that can substantially speed up BoVW and subwindows search based on multi-scales SIFT. We have shown that multi-scales SIFT descriptors are robust for detecting objects with variant sizes. Based on this observation, we have proposed a new coarse-to-fine sub-windows search, which using multi-scales SIFT for reducing search space efficiently. This method can significantly speed up detection without trading off detection accuracy, even better than the traditional method. Our experiments on The GTSDB benchmark have shown the outperform performance of our proposed method.

Acknowledgement. This research is funded by Multimedia Communications Lab, University of Information Technology, VNU-HCM.

References

1. Piccioli, G., De Micheli, E., Campani, M.: A robust method for road sign detection and recognition. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 800, pp. 493–500. Springer, Heidelberg (1994)

2. Barnesi, N., Loy, G., Shaw, D., Robles-Kelly, A.: Regular polygon detection. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 778–785. IEEE (2005)
3. Ruta, A., Li, Y., Liu, X.: Towards real-time traffic sign recognition by class-specific discriminative features. In: BMVC, pp. 1–10 (2007)
4. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, vol. 1, p. 22 (2004)
5. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1615–1630 (2005)
6. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73, 213–238 (2007)
7. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Beyond sliding windows: Object localization by efficient subwindow search. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
8. Zelnik-Manor, L.: On sifts and their scales. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1522–1528. IEEE Computer Society (2012)
9. Bosch, A., Zisserman, A., Muoz, X.: Image classification using random forests and ferns. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
10. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In: International Joint Conference on Neural Networks (submitted, 2013)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
12. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2903–2910. IEEE (2012)
13. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC, vol. 2, p. 5 (2009)
14. Aldavert, D., Ramisa, A., de Mantaras, R.L., Toledo, R.: Fast and robust object segmentation with the integral linear classifier. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1046–1053. IEEE (2010)

Image Feature Extraction and Similarity Evaluation Using Kernels for Higher-Order Local Autocorrelation

Keisuke Kameyama[1] and Trung Nguyen Bao Phan[2]

[1] Faculty of Engineering, Information and Systems, University of Tsukuba
Keisuke.Kameyama@cs.tsukuba.ac.jp
[2] Graduate School of Systems and Information Engineering, University of Tsukuba,
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Abstract. The Higher-Order Moment (HOM) kernel is known to enable efficient utilization of higher-order autocorrelation (HOA) features in signals and images. Several authors report that kernel-based classification methods employing this kernel can classify image textures utilizing the HOA features efficiently. This work evaluates the nature of the HOM kernel of various orders as measures for image similarity. Through sensitivity evaluation and texture classification experiments, it was found that the Local Higher Order Moment (LHOM) kernel enables to control the selectivity of the similarity evaluation by using the Gaussian window.

Keywords: Autocorrelation, Higher-Order Statistics, Kernel.

1 Introduction

In sensing and media retrieval, there is an ever growing need for classification and retrieval of signals, especially images. Traditionally, local Fourier spectra of image signals have been used for characterizing edge directions and textures by way of local autocorrelations and wavelet filters. The generalization of the local Fourier spectra of an arbitrary degree are known as higher-order moment (correlation) functions. Also, their Fourier transforms namely the higher-order spectra such as bispectra, trispectra and so on have been used. These image features are statistics regarding the high order correlation of signal values at certain spatial relations, and those of the phase of multiple frequency components in the signals. Especially, the phase relation between the multiple harmonics components give rich delineation of the image signal structure, and the higher-order spectra (HOS) are known to be useful means for this task [13]. The use of the 4th-order cumulants as a measure of non-Gaussianity in independent component analysis (ICA) is a well known use of higher-order statistics of signals [5].

In recent years, the feasibility of using higher-order statistic feature in pattern recognition are being reexamined under a new light as faster computers become readily available. One such example is the use of high-order local autocorrelation

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 442–449, 2013.

(HLAC) for object recognition and anomaly detection [9][8][12]. However, even in HLAC, the detection of correlations up to 3 points are somewhat limited within small spatial windows, presumably due to the computational demands. The role of higher-order statistics in image signal characterization seems to have not been fully examined yet.

One way to avoid the curse of dimensionality is to use kernel functions instead of directly estimating high dimensional features. As for the higher-order statistics, it has been shown by MacLaughlin et al.[10] that the inner product of higher-order moment (HOM) (higher-order autocorrelation (HOA)) of signals can be calculated with a computation of order $O(k)$, regardless of the order of the moment where k is the number of samples in the digital signal. In recent literature, the HOM kernel function has been used in kernel-based classification methods for signal and image classification [14][4][6].

The author's group extended the HOM kernel to Local Higher-Order Moment (LHOM) kernel, which is a direct generalization of the Gabor filtering extracting the local 2nd order feature to an arbitrary order [7]. We have also shown that kernel functions corresponding to local higher-order moment spectra features (LHOMS kernel) are equivalent to the LHOM kernel [7]. However, the actual conditions for using the kernel as the feature extractor for digital images in a computer, namely how to use the discretized version of the kernel has not been made clear. This work aims to investigate the properties of the HOM kernel when used as a means for feature extraction from digital images.

2 Higher-Order Moment Kernels

2.1 Signal Moments and Higher-Order Moment Kernels

Let $s(\boldsymbol{t})$ be a real valued image signal defined on $\boldsymbol{R}^2$. The n-th order moment (or the $n-1$ order autocorrelation) [13] of $s(\boldsymbol{t})$ is defined as

$$m_{s,n}(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{n-1}) = \int_{\boldsymbol{R}^2} s(\boldsymbol{t}) \prod_{k=1}^{n-1} s(\boldsymbol{t} + \boldsymbol{\tau}_k) d\boldsymbol{t}. \tag{1}$$

Traditionally, moments have been used for characterization of image and texture. However, direct calculation of high order moments suffer from the required computation, and moments of order $n > 3$ have been rarely used for image characterization. Image features of orders up to 3 within a limited spatial shifts ($\boldsymbol{\tau}_k$ in Eq. (1)) specifically known as higher-order local autocorrelation (HLAC) have been successfully applied to image and video classification [9][8][12].

In [10], McLaughlin and Raghu showed that inner product of moment functions of arbitrary order can be calculated without a drastic increase of computation. They showed that the inner product of the n-th order moments of two signals s and v can be calculated as,

$$K_n(s, v) = \langle m_{s,n}, m_{v,n} \rangle = \int \left[\int s(\boldsymbol{z}) v(\boldsymbol{z} + \boldsymbol{\tau}) d\boldsymbol{z} \right]^n d\boldsymbol{\tau}. \tag{2}$$

In modern terms of pattern recognition, this inner product is the kernel function of n-th order moment feature extractor, and enables a tractable use of higher-order moment features in the frameworks of kernel-based pattern recognition methods. In the following, the two-variable function $K_n(s,v)$ in Eq. (2) will be referred to as the higher-order moment (HOM) kernel function.

2.2 Local Signal Moments and Local Higher-Order Moment Kernels

Because the HOM kernel is based on the nonlocalized moment function of Eq. (1), it can work as a similarity measure of HOM feature for the whole image. In contrast, when the characterization needs to be limited to a local image portion, the local higher-order moment kernel (LHOM kernel)

$$K_{w,n}(s,v\ ;\ \boldsymbol{x},\boldsymbol{y}) = \int \left[\int w(\boldsymbol{z}) s(\boldsymbol{z}+\boldsymbol{x}) w(\boldsymbol{z}+\boldsymbol{\tau}) v(\boldsymbol{z}+\boldsymbol{y}+\boldsymbol{\tau}) d\boldsymbol{z} \right]^n d\boldsymbol{\tau}. \qquad (3)$$

can be used [7]. Here, function $w(\boldsymbol{z})$ is the spatial window centered at positions $\boldsymbol{x}$ and $\boldsymbol{y}$ of signals s and v, respectively. Thus, the LHOM kernel will enable to evaluate the similarity of local higher-order moments.

In [7] and [6], the equivalence of the HOM kernel to the inner product of higher-order moment spectra (HOMS kernel), and the equivalence of the LHOM kernel to the inner product of local higher-order moment spectra (LHOMS kernel) have been shown. Accordingly, evaluation of the HOM kernel for $n = 2$ amounts to a comparison of the power spectra of two images. Similarly, the LHOM kernel for $n = 2$ will evaluate the similarity of the local power spectra that may be extracted using an array of 2 dimensional Gabor filters [3][1]. HOM kernel of $n = 3$ and $n = 4$, are equivalent to inner products of bispectra and trispectra of the signals, respectively. Spectral features of $n > 2$ are capable of characterizing the phase of the harmonic frequency components in the signal, to which power spectral features ($n = 2$) are insensitive.

2.3 Discrete Calculation of HOM Kernels

When HOM and LHOM kernels are to be used for feature extraction from digital images, the discrete version of the kernels

$$K'_n(s,v) = \sum_i \left[\sum_j s(\boldsymbol{z}_j) v(\boldsymbol{z}_j + \boldsymbol{\tau}_i) \right]^n \qquad (4)$$

and

$$K'_{w,n}(s,v\ ;\ \boldsymbol{x},\boldsymbol{y}) = \sum_i \left[\sum_j w(\boldsymbol{z}_j) s(\boldsymbol{z}_j+\boldsymbol{x}) w(\boldsymbol{z}_j+\boldsymbol{\tau}_i) v(\boldsymbol{z}_j+\boldsymbol{y}+\boldsymbol{\tau}_i) \right]^n \qquad (5)$$

are to be used. Although they have been successfully used in support vector machines for texture classification [7] [6], close investigation regarding their properties in characterizing the image similarities are yet to be made. This paper aims to delineate the nature of the HOM and LHOM kernels when they are used as similarity measures of images.

3 Sensitivity Analysis

When HOM and LHOM kernels are to be used in the kernel-based classification algorithms, the kernel will function as a measure of similarity between (L)HOM features of two signals. In order to shed light to kernel function used as similarities, here we will experimentally evaluate the sensitivity of the kernels against small discrepancies in the images. We will focus on the difference of order and the use of window functions (HOM or LHOM).

The kernel values $K'(\boldsymbol{x}_0, \boldsymbol{x}_m)$ for two input images, namely a *standard* image $\boldsymbol{x}_0$ and a *modified* image $\boldsymbol{x}_m$ $(m = 1, 2, \ldots)$ will be observed as the modification is gradually made significant. The changes applied to $\boldsymbol{x}_0$ to make $\boldsymbol{x}_m$ will be scaling, rotation and noise addition that can be common in image matching scenarios. Through this investigation, it is aimed that the relative sensitivity to the above fluctuations will become clear.

Kernels

HOM kernels and LHOM kernels of Eqs. 4 and 5 of orders $n = 2, 3, 4, 5$ were used. A normalized Gaussian window $w(\boldsymbol{t}, \Sigma) = (2\pi)^{-1}|\Sigma|^{-1/2} \exp(-0.5\boldsymbol{t}^T \Sigma^{-1}\boldsymbol{t})$ with $\Sigma = diag(\sigma^2)$ was used for LHOM kernels, Parameter σ was chosen to be 1/3 of the image width.

Image Data

A natural image of grass leaves shown in Fig. 1(a) was chosen from the Vision Texture collection [11]. As this image has much randomness by nature, it has a broad distribution in moment and spectral feature space.

Image modifications

The original image was set to be the *standard* images mentioned in the previous section. The changes in the *modified* images are,

1. Scaling (ratio $\in \{1, 1, 01, 1.03, 1.05, 1.1, 1.15\}$)
2. Rotation (angle $\in \{0, 1, 3, 5, 10, 20\}$ (degree))
3. Noise addition (SNR $\in \{\infty, 50, 30, 20, 10, 0\}$)

Bicubic interpolation was used upon scaling and rotation of natural images, and Gaussian noise was added to the original images so that the signal-to-noise ratio (SNR) of the image will be at the specified ratio. Examples of the modifications are shown in Fig. 1(b)-(d).

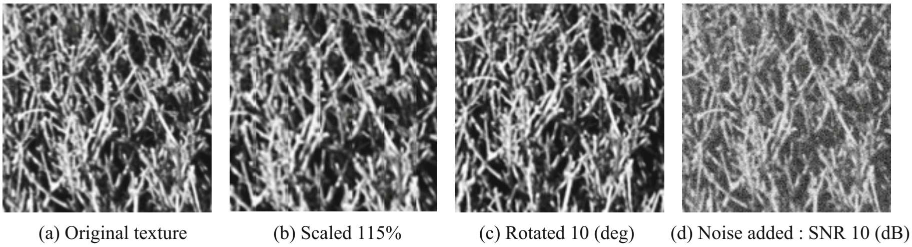

(a) Original texture (b) Scaled 115% (c) Rotated 10 (deg) (d) Noise added : SNR 10 (dB)

Fig. 1. Texture "Leaves 0013" from the Vision Texture collection

Procedure

For each standard image $\boldsymbol{x}_0$ and a set of modified images $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, relative kernel response defined as

$$r_m = \frac{K'(\boldsymbol{x}_m, \boldsymbol{x}_0)}{K'(\boldsymbol{x}_0, \boldsymbol{x}_0)} \quad (m = 1, 2, \ldots) \tag{6}$$

were evaluated for each type of modifications.

Results and Comments

In Fig. 2, the relative HOM and LHOM kernel responses for the original and modified images are shown.

Scaling and Rotation: Generally, the selectivity of the kernels increased with the order. Kernels of odd order have significantly higher selectivity in HOM kernels when compared with those of even orders. However, this nature was not observed in the LHOM kernels. As the Gaussian window used in LHOM kernels functions as the *blurring* filter in the spectral domain of signals, it can be considered that it reduces the sparseness within the high dimensional feature space. This sparseness may be controlled by choosing the width of the Gaussian windows as in Gabor filters [3]. However a deeper investigation on this nature of LHOM kernels is due.

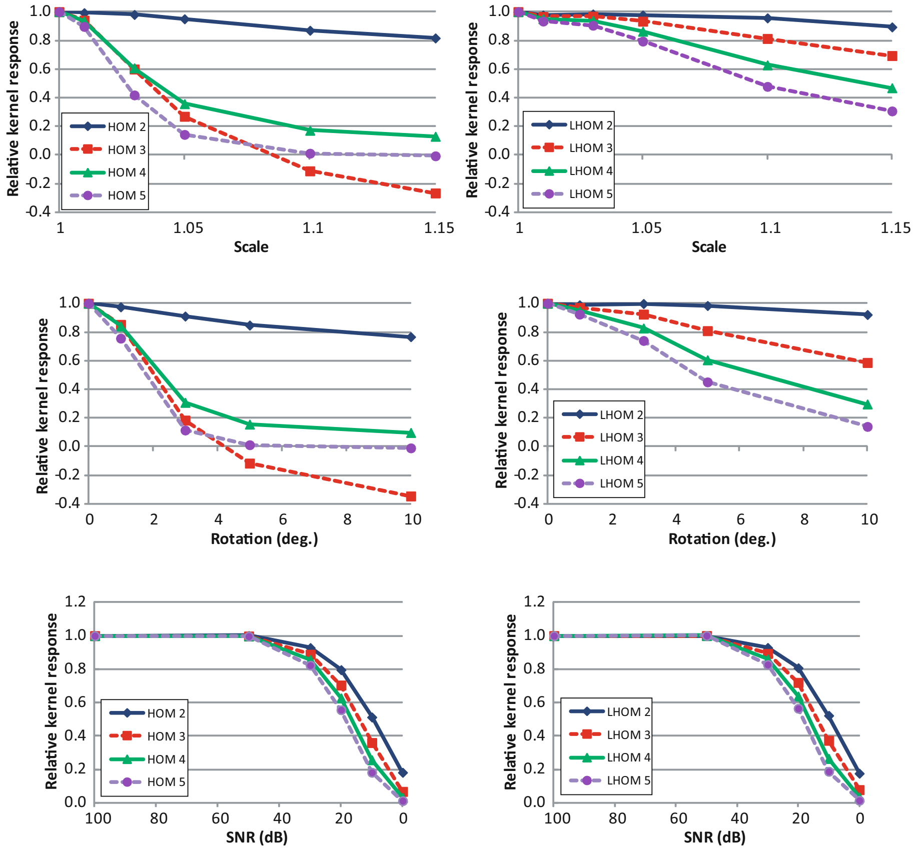
Noise Addition : Relative kernel responses dropped for all kernels as the noise in the modified images were increased. Both kernels of higher order dropped faster, and no differences between HOM and LHOM kernels were found.

4 Texture Classification

HOM and LHOM kernels of various order were applied to texture classification problems using Support Vector Machines (SVM). This experiment aims to evaluate the basic nature of HOM and LHOM kernels observed in the sensitivity analysis, in image classification problems.

Image Set

10 natural texture classes included in the Vision Texture dataset [11] were chosen. From each image, 20 training and 20 test subimages were cut out from random positions. The size of the subimages were 31 × 31 and 64 × 64 pixels.



(a)

(b)

Fig. 2. Relative responses of (a)HOM and (b)LHOM kernels for changes in scale, rotation and additive noise level

Kernels and SVM

HOM and LHOM kernels of orders $(n = 2, \ldots, 9)$ were used. Soft margin SVM with regularization parameter of $(C = 100)$ by the libSVM toolkit [2] was used. For the multiclass problem, one-against-others strategy was employed.

Conditions

Each SVM was repeatedly trained and tested by different sets of training and testing images. Average test rate of 10 trials were used for evaluation.

Results and Comments

In Fig. 3, the average classification rates for the test set are shown for two different image sizes. For both image sizes, it is clear that the rates when LHOM kernels were used gave superior results in comparison with HOM kernels. This difference may be explained from the results of the sensitivity analysis. Also

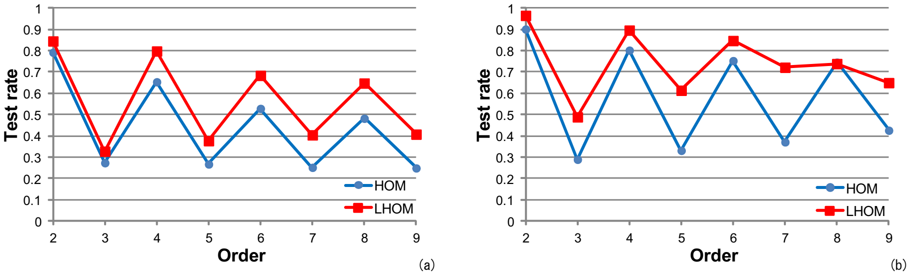


Fig. 3. Classification rates for SVMs with HOM and LHOM kernels of orders 2 to 9. Train/test image size : (a) 31 × 31 pixel (b) 64 × 64 pixel.

notable is the tendency that even order kernels tend to perform better than odd order kernels. Similar results have been reported in [4], and we conjecture that they are due to the higher sensitivity in odd orders. This tendency is somewhat relaxed when LHOM kernels were used, and it may be controlled by the widths of the window function.

In this experiment, the high sensitivity of HOM kernels (especially in higher orders) seems to be negatively contributing. However, it may turn out to be a plus when subtle differences in the image become key to correct classification, such as in biometric authentication problems.

5 Conclusion

This report explored the nature of Higher-Order Moment (HOM) kernel and Local HOM kernel functions to be used as feature extractors in the kernel-based pattern recognition methods. The sensitivities of the kernels against image changes have been investigated, and application to SVM's in texture classification problems were reported. It was found that the selectivity of the kernels increase with the its moment order, and it was discussed that the the use of window functions in LHOM kernels may contribute to control the sensitivity. This feature can be useful in applying the kernel to problems having different sparsity in the feature space.

Acknowledgment. This work was supported by Grant-in-Aid for Scientific Research 22500198 by Japan Society for the Promotion of Science (JSPS).

References

1. Bovik, A.C., Clark, M., Geisler, W.S.: Multichannel texture analysis using localized spatial filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(1), 55–73 (1990)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27 (2011)

3. Daugman, J.G.: Uncertainty relation for resolution in space, spatial frequency and orientation optimized by two-dimensional visual cortical filters. Journal of the Optical Society of America A 2(7), 1160–1169 (1985)
4. Horikawa, Y.: Comparison of support vector machines with autocorrelation kernels for invariant texture classification. In: Proc. 17th International Conference on Pattern Recognition (ICPR 2004), vol. 1, pp. 660–663 (2004)
5. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley-Interscience (2001)
6. Kameyama, K.: Comparison of local higher-order moment kernel and conventional kernels in SVM for texture classification. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part I. LNCS, vol. 4984, pp. 851–860. Springer, Heidelberg (2008)
7. Kameyama, K., Taga, K.: Texture classification by support vector machines with kernels for higher-order gabor filtering. In: Proceedings of International Joint Conference on Neural Networks 2004, vol. 4, pp. 3009–3014 (2004)
8. Kobayashi, T., Otsu, N.: Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation. In: Proceedings of International Conference on Pattern Recognition, pp. 741–744 (2004)
9. Kurita, T., Otsu, N.: A new scheme for practical, flexible and intelligent vision systems. In: Proceedings of IAPR Workshop on Computer Vision, pp. 431–435 (1988)
10. MacLaughlin, J.A., Raviv, J.: N-th autocorrelations in pattern recognition. Information and Control 12(2), 121–142 (1968)
11. MIT Vision and Modeling Group: Vision texture (1995)
12. Nanri, T., Otsu, N.: Unsupervised abnormality detection in video surveillance. In: Proceedings of MVA 2005 (2005)
13. Nikias, C.L., Petropulu, A.P.: Higher-Order Spectra Analysis - A Nonlinear Signal Processing Framework. Prentice Hall (1993)
14. Popovici, V., Thiran, J.P.: Higher order autocorrelations for pattern classification. In: Proceedings of the International Conference on Image Processing 2001, pp. 724–727 (2001)

Integrated Multi-scale Retinex Using Fuzzy Connectivity Based on CIELAB Color Space for Preserving Color

Biho Kim, Wonyong Jo, and Hyung-Min Park

Department of Electronic Engineering, Sogang University,
35 Baekbeom-ro, Mapo-gu, Seoul 121-742, Republic of Korea
{biho,crestig,hpark}@sogang.ac.kr

Abstract. In this paper, a tone reproduction method that preserves color in area containing unevenly distributed intensity is proposed. Although Retinex algorithm is one of the most popular tone reproduction algorithms, it still has several problems caused by using only the Gaussian kernel which is insufficient to consider local condition such as steep variation of illumination or sudden change of color. To overcome this problem, we proposed a pixel-wise adaptive fuzzy kernel which is capable for changing its weight to preserve the local condition in an image, instead of the Gaussian kernel. Moreover, the proposed kernel also considered color consistency as human visual system by eliminating illumination components in an observed image. Experiment results on some images with different local conditions showed that the proposed method effectively reduces the problems.

Keywords: Tone reproduction, multi-scale Retinex, fuzzy connectivity.

1 Introduction

Although camera has been continually developed, it is still not good enough compare to human visual system. One represent problem is low dynamic range problem due to physical limitation of camera sensor. The image containing both low and high lightness regions cannot be represented well by using present day technology of sensor. In contrast, human eye has high dynamic range because human visual system can be locally adapted for lightness variation. Currently, various algorithms are trying to imitate this functionality of human visual system.

There are several methods based on Retinex theory to solve low dynamic range problem [1-3]. These algorithms modeled an image as multiplication of reflectance component and illuminance component, and performed tone reproduction by eliminating illuminance component from image. Multi-scale Retinex (MSR) algorithm is recently proposed and that is a weighted sum of Single-scale Retinex (SSR) with Gaussian kernel of various scales [4-6]. Defects such as halo effect which was caused in SSR could be complemented by using the scales. However, ratio of RGB channel is distorted by independent process in MSR and Integrated Multi-scale Retinex (IMSR) was proposed to resolve the problem [7]. In this algorithm, surround image was computed on a single channel, and applied to every RGB channel preserving the RGB

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 450–457, 2013.

ratio. Even with the effort, human visual system still perceives some distortions in the result image. IMSR based on CIELAB color space algorithm was proposed to reduce this defect [8]. This approach calculates surround image using only L^* channel, and enhances the L^* channel using the surround image. However, using only L^* channel can cause low saturation problem and it makes saturation compensation very essential to enhance it.

The most of Retinex algorithms mentioned above use the Gaussian kernel to calculate surround image. Even they have good performance results for general images, they are still improper for real world image which has various problems. For example, if there is inconsistent illumination area in local range, assumption made in Gaussian kernel, illumination is same in a local range, is not suitable. Furthermore, since image in CIELAB color space is calculated from RGB value, it makes the problem to calculate L^* color space even if pixel color is different in same lightness condition. In other words, the difference of L^* values makes different surround image even with the same illumination and this causes error to calculate surround image.

Inspired by this point, we propose to modify Retinex algorithm to get more accurate surround image. This paper proposes to use fuzzy connectivity kernel to improve visibility while reducing color distortion in inconsistent illumination area or inconsistent color area. Fuzzy connectivity kernel is defined in each pixel, based on a^* and b^* channel information. It makes adaptive surround image corresponding to each pixel and its neighboring pixels, and it helps to perform better.

In the rest of the paper is organized as follows: Section 2 briefly describes conventional tone reproduction algorithm. Section 3, the proposed algorithm is presented. In section 4, experimental results are presented and Section 5 describes some concluding remarks.

2 Previous Retinex Method

2.1 Integrated Multi-scale Retinex

Conventional MSR method used RGB channel to generate surround image and it is applied to each RGB channel independently. This changes RGB ratio in a pixel and causes color distortion. IMSR method proposed by Wang only used luminance channel to form surround image to fix the problem [7]. The surround images are generated by using Gaussian kernels with different standard deviations, and a weighted sum of these images forms an integrated surround image. It is applied to each color channel to keep the color balance. The result image of IMSR method is given by

$$SSR_i(x, y) = A \frac{I_i(x, y)}{S_{sum}(x, y)}, \tag{1}$$

where I is the input RGB image, i is the RGB channel index, A is a gain coefficient, (x, y) is position of a pixel, and S_{sum} is integrated surround image.

Integrated surround image in Eq. (1) is computed by

$$S_{sum}(x,y)=\sum_{m=1}^{M} w_m S_m(x,y,\sigma_m)\,, \tag{2}$$

S_{sum} denotes weighted sum of surround images from different scales, M is number of scale, σ_m is a standard deviation of surround image and w_m is weight of surround image $S_m(x,y,\sigma_m)$. Each surround image is convolution of luminance Y and Gaussian kernel $G_m(x,y,\sigma_m)$ employing different standard deviation as follows

$$S_m(x,y,\sigma_m)=G_m(x,y,\sigma_m)*Y(x,y)\,. \tag{3}$$

2.2 Integrated multi-scale Retinex based on CIELAB Color Space

Many Retinex methods including IMSR are taking place in RGB color space. Hue distortion in CIELAB color space which indicates hue distortion in human visual system can be raised by these methods.

To prevent these problems, IMSR based on CIELAB color space takes place in device-independent color space, CIELAB [8]. In this method, input RGB image is transformed to CIELAB color space, and IMSR is applied to only L^* channel to preserve balance of colors components. This enhances luminance of image, but not saturation, therefore the dark regions in the original image have low saturation values after Retinex method. Saturation adjustment need to be performed on a^*, b^* channel according to the luminance change to enhance the saturation, because of the unnatural saturation problem in this process. After saturation adjustment, the adjusted CIELAB image is transformed back to RGB color space. Luminance enhanced image is given by applying IMSR to only L^* channel and it is acquired by

$$L_{sum}{}^*(x,y)=A\frac{L^*(x,y)}{S_{sum}(x,y)}\,, \tag{4}$$

where

$$S_{sum}(x,y)=\sum_{m=1}^{M} w_m S_m(x,y,\sigma_m)\,, \tag{5}$$

where

$$S_m(x,y,\sigma_m)=G_m(x,y,\sigma_m)*L^*(x,y)\,. \tag{6}$$

In Eq. (4) and (6), L^* is the luminance component of an image, and the others are same as IMSR process.

Normalization is needed to keep the values in displayable range. Instead of using maximum luminance value, the value having zero gradients in luminance cdf is employed because the number of maximum luminance value is very small and they can be considered as noise.

As mentioned before, saturation adjustment is also needed. The saturation adjustment in proportion to luminance enhancement is performed by following equation.

$$C_{adj}(x,y) = C_{in}(x,y)\frac{GB(L^{*}{}_{adj})}{GB(L^{*})}. \tag{7}$$

In Eq. (7), C_{in} is a chroma value of an original image, C_{adj} is an adjusted chroma value, and $GB(L^*)$ is sRGB gamut boundary corresponding to luminance. $GB(L^*{}_{adj})$ is modified gamut boundary by enhanced luminance. In the saturation enhancement process, oversaturation problem is solved by using the ratio of sRGB gamut boundary.

3 Proposed Method

Previous methods used Gaussian kernel to acquire surround image in L^* channel to keep the balance of colors components [8]. However, using only L^* channel also has a problem. If there are two pixels under same illumination, L^* value can be different if the colors of the pixels are different because L^* value is computed from RGB value. Therefore, it is not easy to obtain accurate surround image in the color inconsistent area even with the same illuminance. Furthermore, if same color pixels are under different illumination condition, each of them must be enhanced differently but Gaussian kernel enhances with same amount because it assumes that illumination of local area is changing slowly.

To improve this color distortion problem, we propose that a kernel with fuzzy connectivity to use not only the spatial distance like Gaussian kernel but also the color difference.

$$K(x,y,\sigma_m) = \frac{G_m(x,y,\sigma_m)}{1+k\left\|f(x,y)-f(c,d)\right\|}, \tag{8}$$

where $G_m(x,y,\sigma_m)$ is a Gaussian kernel whose position of center pixel is (x,y), σ_m is a standard deviation of kernel, (c,d) means position of pixel except (x,y) in Gaussian kernel $G_m(x,y,\sigma_m)$, f is two dimensional vector which means color information of a pixel in a^*, b^* space, and k is a parameter determining how much

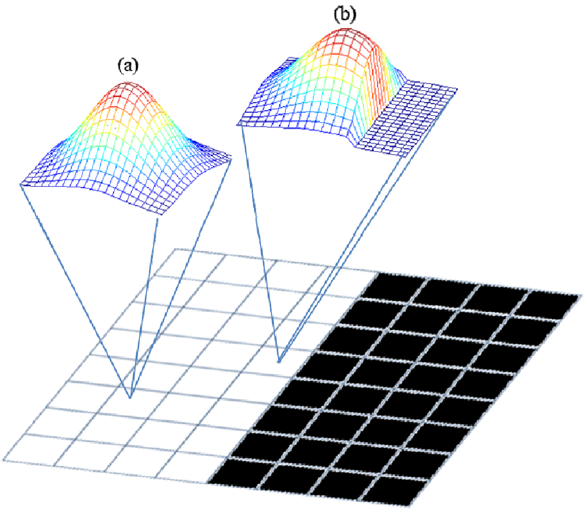


Fig. 1. Representation of proposed kernel

color information is used to yield kernel weight. Instead of Gaussian kernel used in Eq. (6), the new kernel $K(x,y,\sigma_m)$ is utilized in proposed method to make surround image as follows,

$$S_m(x,y,\sigma_m) = K(x,y,\sigma_m) * L^*(x,y)\,. \tag{9}$$

Gaussian kernel was identically applied to every pixel of image in previous methods, but we used adaptively generated kernel $K(x,y,\sigma_m)$ for each pixel using color difference between (x,y) and its surrounding pixels to reduce color distortion. Fig .1 depicts that proposed kernel is applied to the original image in order to enhance luminance. In Fig. 1, a white pixel determining the kernel (a), is surrounded by other white pixels therefore $f(x,y)$ and $f(c,d)$ become identical values. It makes the denominator in Eq. (8) ignorable. However, the right side of another white pixel kernel (b) is almost zero due to the large color difference between center pixel and black neighboring pixels.

4 Experiment

The proposed method was evaluated using image which is taken by Canon IXUS 110 IS digital camera supporting sRGB profile and some well-known images which are generally used for Retinex algorithm tests.

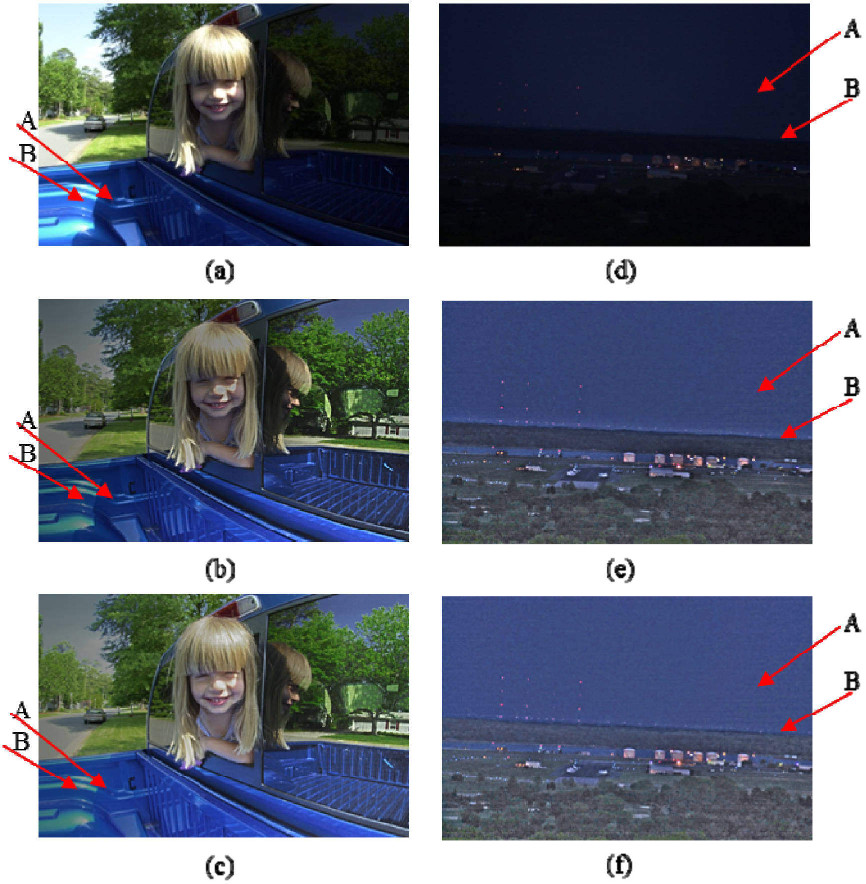


Fig. 2. Results on general image. (a), (d) : Original image, (b), (e) : IMSR based on CIELAB, (c), (f) : Proposed.

The first experiment is performed about general test images. In Fig. 2(a), point A and B are same color, but in different illuminance condition. Even though the point A should be enhanced more than B to eliminate effect of illuminance, Gaussian kernel makes same amount of enhancement for both A and B in Fig. 2(b). In Fig. 2(c), luminance difference between point A and B is smaller than Fig. 2(b), because there is more enhancement in point A by using proposed kernel. On the other hand, In Fig. 2(d), point A and B are originally very similar but in fig 2(e) these two points seems to be different. Since the points have dissimilar neighboring pixels, amount of enhancement is different for points A and B. With suggested algorithm, we can verify this difference is reduced in Fig. 2(f).

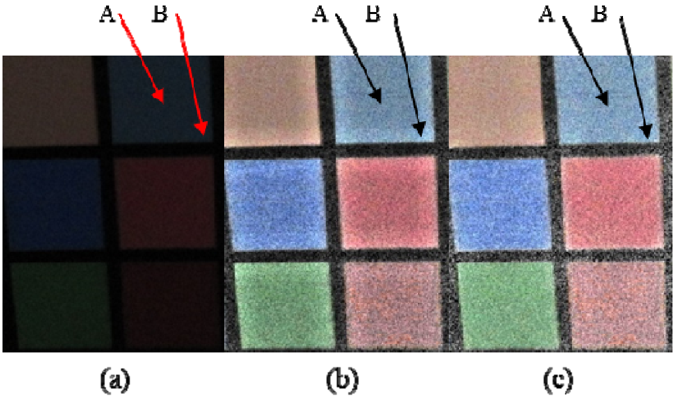


Fig. 3. Results on color checker image. (a) : original image, (b) : IMSR based on CIELAB, (c) : proposed.

In the second experiment to confirm the proposed method reduces color distortion more than previous method, color checker image was used. In Fig. 3(a), illumination condition and color is constant in point A and B, therefore these two points are same. However, point A is seen darker than B in Fig. 3(b) because neighboring pixels of A and B are different. In Fig. 3(c), we can verify with the eye that the color difference between point A and B is smaller than the Fig. 3(b). To confirm more quantitatively, we conversed images of Fig. 3(a) and (b) to HSV color space then compared standard deviation of Hue. The Hue value of each color regions must be same and it makes smaller standard deviation of Hue represents smaller color distortion. The test image has 7 color regions, therefore we were able to model these two test images by Gaussian Mixture Model with 7 components, and find standard deviation of each component. Then, the sum of 7 standard deviations can be considered as standard deviation of Hue. As a result, Fig. 3(c) has smaller standard deviation of Hue than Fig. 3(b) which means degree of color distortion is reduced in proposed method than previous method. This result is depicted in Table 1.

Table 1. Sum of standard deviation of Hue in HSV color space

Method	CIELAB IMSR	Proposed method
Sum of standard deviation of Hue	8945	3299

5 Conclusion

In this paper, we presented a tone reproduction method that preserves color in area containing unevenly distributed intensity. Conventional Retinex algorithm performed in CIELAB color space has unwanted distortion in the area including steep variation of illumination or sudden change of color. To prevent this distortion while acquiring surround image from L^* channel, we proposed an adaptive kernel using fuzzy connectivity based on the color information instead of using the Gaussian kernel only. We can preserve local color information by using this kernel while the Gaussian kernel only takes integrated value of it. The experiment results showed the improved visibility of low dynamic range images. More precisely, result of proposed algorithm showed better local consistency when observed image contained sudden changes of intensity compared to conventional algorithm.

References

1. Rahman, Z.: Properties of a center/surround Retinex: Part 1: Signal processing design. NASA Contractor Report 198194 (1995)
2. Jobson, D.J., Rahman, Z., Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE Trans. Image Processing 6 (1997)
3. Land, E.H.: An alternative technique for the computation of the designator in the Retinex theory of color vision. Proc. Natl. Acad. Sci. U.S.A. 83, 3076 (1986)
4. Rahman, Z., Jobson, D., Woodell, G.A.: Multiscale retinex for color image enhancement. In: Proc. IEEE International Conference on Image Processing. IEEE (1996)
5. Jobson, D.J., Rahman, Z.: A Multiscale Retinex for Bridging the Gap Between Color Images and the Human Observation of Scenes. IEEE Trans. Image Processing 6 (1997)
6. Rahman, Z., Jobson, D.J., Woodell, G.A.: Multiscale Retinex for color rendition and dynamic range compression. In: Proc. SPIE, vol. 2847, p. 183 (1996)
7. Wang, L., Horiuchi, T., Kotera, H.: High Dynamic Range Image Compression by Fast Integrated Surround Retinex Model. J. Image Science and Technology 51(1) (2007)
8. Kyung, W.-J., Lee, T.-H.: Improved color reproduction based on CIELAB color space in integrated multi-scale retinex. In: Proc. SPIE Electronic Imaging, vol. 7241 (2009)

Object Pose Estimation by Locally Linearly Embedded Regression

Bisser Raytchev*, Kazuya Terakado, Toru Tamaki, and Kazufumi Kaneda

Department of Information Engineering, Hiroshima University, Japan
{bisser,terakado,tamaki,kin}@hiroshima-u.ac.jp

Abstract. In this paper we propose a new local learning algorithm for appearance-based object pose estimation, called Locally Linearly Embedded Regression (LLER). LLER uses a constrained version of Locally Linear Embedding (LLE) to simultaneously embed into an intermediate low-dimensional space the training images, the query image and a grid of pose parameters. A linear map is learned between the points in the local neighborhood of the query representation in this low-dimensional intermediate space and their corresponding pose parameters, which is used to directly recover the pose of the query image. The proposed method has been evaluated in a pose estimation task on a database of 16 different objects, consistently outperforming several representative global and local appearance-based pose estimation methods.

1 Introduction

The estimation of the pose of a 3D object from a single 2D image is one of the most important problems in computer vision, with numerous applications in natural human-computer interfaces, robotic vision, augmented reality and so on. Most of the methods for pose estimation can be classified into two major groups: *model-based* and *appearance-based* (or view-based) approaches. Model-based approaches typically proceed by matching features extracted from the query image to a pre-built 3D model of the same object. Some general representative methods include the hypothesize-and-test method [1], geometric hashing [2], pose clustering [3], etc., see e.g. [4] for a survey. Appearance-based approaches [5,6,7,8] typically learn a linear or non-linear map between the available training images and their corresponding pose parameters, which is then used to recover the pose of a test query image by directly mapping it to pose space. Both approaches have their strengths and limitatations, but in this paper we follow the appearance-based approach's point of view to pose estimation.

Most appearance-based methods learn a single *global* map between all training samples and their corresponding pose parameters. However, representing the relation between all views of an object and the pose parameters through a single map makes the problem unnecessarily complicated. A more simple and efficient solution is to use *local learning*, where the map is learned only for a small subset

* Corresponding author.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 458–468, 2013.

of the training images, which is in the local neighborhood of the query image. In this case even a simple linear map would be adequate.

In this paper we propose a new appearance-based local learning algorithm, called Locally Linearly Embedded Regression (LLER), which uses a constrained version of Locally Linear Embedding (LLE) [9,10] to simultaneously embed into an intermediate low-dimensional space both the training images, the query image and a grid of pose parameters. Then a linear map is learned between the points in the local neighborhood of the query representation in this low-dimensional intermediate space and their corresponding pose parameters, which is used to directly recover the pose of the query image. We have evaluated the proposed method in a pose estimation task on a database of 16 different objects. The results show that our method consistently outperforms several representative global and local appearance-based pose estimation methods.

In the next section we first briefly review relevant related work and also describe the motivation behind the design of our new local learning method, which is introduced in Section 3. Section 4 describes experimental results, and section 5 then concludes the paper.

2 Related Work

Our work is most strongly related to [11] who proposed an appearance-based pose estimation method, called Local Procrustes Regression (LPR), which uses local learning to estimate the unknown pose of an object from a single query or test image. LPR first finds the k-nearest neighbors of the test image $\boldsymbol{x}_T$ among all available training images $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$. The corresponding pose parameters for the training images in the k-neighborhood of the test image are $\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k$. First LPR uses Multidimensional Scaling (MDS) [12,13] to embed together the test image and the k training images into a low-dimensional space, where their representations are $\boldsymbol{y}_T$ and $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k$, respectively. Since the correspondence between the training images $\boldsymbol{x}_i$ and their pose parameters $\boldsymbol{p}_i$ is known, and therefore the correspondence between $\boldsymbol{y}_i$ and $\boldsymbol{p}_i$ is also known, the transformation between them is calculated to minimize the cost

$$R = \sum_{i=1}^{k} (\boldsymbol{p}_i - s\boldsymbol{A}^T \boldsymbol{y}_i - \boldsymbol{b})^T (\boldsymbol{p}_i - s\boldsymbol{A}^T \boldsymbol{y}_i - \boldsymbol{b}) \tag{1}$$

where the orthonormal matrix $\boldsymbol{A}$ represents the rotation/reflection, $\boldsymbol{b}$ the translation, and s the isotropic scaling, needed to align the low-dimensional representations $\boldsymbol{y}_i$ and their corresponding pose parameters $\boldsymbol{p}_i$. The cost in Equation (1) minimizes the "goodness of fit" criterion [14] and in essence finds the *similarity transformation* which optimally aligns the data embedding and the parameters.

Although in [11] it is shown that on a pose estimation task including 16 different objects, LPR generally outperforms global linear regression and achieves similar performance to Support Vector Regression (SVR), still there are some drawbacks which need to be addressed. First, the use of the similarity transformation in Eq. 1 is unnecesary limiting, as it is unlikely that the relation between the embedded data and the corresponding parameters (even if only the data in a small local neighborhood is considered) can be well-represented by a similarity transformation (rotation, translation and isotropic scaling). A better choice would be to use a more general linear transformation, like the *affine transformation* with non-isotropic scaling. Then the linear map A between the low-dimensional embedding Y_k of the k-neighbors of the test image and their corresponding pose parameters P_k can be found as

$$A = Y_k^+ P_k \tag{2}$$

where Y_k^+ is the Moore-Penrose generalized inverse matrix [15] of Y_k. Then the pose of the test image can be calculated as

$$p_T = y_T A = y_T Y_k^+ P_k. \tag{3}$$

We will call this local learning method based on affine transformation Local Affine Regression (LAR) and will compare it with LPR and LLER in section 4.

Another problem with LPR is that the low-dimensional embedding of the images is obtained through MDS, which is a *linear* method (also known as Principal Coordinates Analysis [14], and equivalent to PCA [16], although obtained from the distance matrix of the data, rather than the covariance matrix). Recently, manifold learning-based *non-linear* methods for locality preserving dimensionality reduction [9,17,18,19] have been shown to be able to represent more accurately the low-dimensional manifold structure of object images, although their use in the framework of local learning has not attracted much attention yet. Note that non-linear manifold learning methods can be quite computationally expensive for large-scale datasets, but this is not a problem in the context of local learning, as in this case only the data in the neighborhood of the query/test sample is relevant and needs to be accounted for.

In the following section we propose a new local learning-based pose estimation method, which addresses simultaneously the above-mentioned limitations of LPR.

3 Locally Linearly Embedded Regression (LLER)

In this section we will describe the proposed Locally Linearly Embedded Regression (LLER) method for pose estimation. In order to understand how LLER works, it would be advantageous to know how Locally Linearly Embedding (LLE)[9] works, and especially its constrained version, Constrained Locally Linearly Embedding (cLLE)[10], on which it is based. Therefore, we will first briefly review these methods.

LLE is a manifold-learning method which finds a low-dimensional embedding of the data, while at the same time preserving locally the linear structure of neighboring data. If N d-dimensional data vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ are given, their low-dimensional representation is found by LLE through the following 3 steps:

step 1 For each data point $\boldsymbol{x}_i$ find its k-nearest neighbors $\boldsymbol{x}_j$ $(j = 1, \cdots, k)$, using the Euclidean distance.

step 2 Compute the local reconstruction weights W_{ij} through which $\boldsymbol{x}_i$ is linearly reconstructed from its neighbors by minimizing the reconstruction error below

$$\varepsilon(W) = \sum_i ||\boldsymbol{x}_i - \sum_j W_{ij}\boldsymbol{x}_j||^2 \tag{4}$$

subject to the constraint $\sum_j W_{ij} = 1$. Using Lagrange multiplier to enforce the constraint, the optimal weights are obtained as

$$w_{ij} = \frac{\sum_k G_{jk}^{-1}}{\sum_{lm} G_{lm}^{-1}} \tag{5}$$

where G is the Gram matrix $G_{jk} = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_k)$.

step 3 Using the weights W_{ij} found in step 2, compute the low-dimensional representations $\boldsymbol{y}_i$ of each $\boldsymbol{x}_i$ which minimize

$$\varPhi(Y) = \sum_i ||\boldsymbol{y}_i - \sum_j W_{ij}\boldsymbol{y}_j||^2 \tag{6}$$

subject to the constraint $\langle \boldsymbol{y_i y_i^T} \rangle = I$. LLE minimizes (6) by computing the d eigenvectors of $M = (I - W)(I - W)^T$ corresponding to the smallest nonzero eigenvalues, which put together give the d-dimensional embedding $\boldsymbol{Y}$ of the data.

Next, Constrained LLE (cLLE) will be explained. It has been used in [10] to find a common low-dimensional representation of two different high-dimensional data sets. Assume that we have two high-dimensional data sets: one data set X^1 represented by n_1 d_1-dimensional vectors $\boldsymbol{x}^1$, and a second data set X^2 represented by n_2 d_2-dimensional vectors $\boldsymbol{x}^2$.

It is also assumed that n_c of the data samples are *in correspondence*, i.e. they have some common characteristic (or even could be identical data, the intersection of the two data sets), and we want these corresponding data samples to have the same low-dimensional representation. Note that n_c is smaller than both n_1 or n_2. The two data sets can be represented by the following block matrix:

$$X = \left[X_c \middle| X_s \right] = \left[\begin{array}{c|cc} X_c^1 & X_s^1 & X_r^1 \\ X_c^2 & X_r^2 & X_s^2 \end{array} \right]. \tag{7}$$

In (7), X is a $(d_1 + d_2) \times (n_1 + n_2 - n_c)$ matrix, where the data in correspondence from the two data sets is stored in X_c, occupying the first n_c columns of X. X_s^1 contains the remaining available data x_j^1 $(j = (n_c + 1), \cdots, n_1)$ from the first set, and X_s^2 the remaining available data x_k^2 $(k = (n_c + 1), \cdots, n_2)$ from the second set, which are not in correspondence. X_r^1 and X_r^2 is *missing data*, i.e. X_r^1 is the yet unknown data from the first set corresponding to X_s^2 from the second set, and X_r^2 is the unknown data from the second set corresponding to X_s^1 from the first set,

The task of constrained LLE is to recover the missing data in X. If the two data sets are embedded through LLE without using any constraints, this would result in two separate low-dimensional representations,Y^1 and Y^2, each of which can be obtained by diagonalizing $M^1 = \left(I - W^1\right)\left(I - W^1\right)^T$ and $M^2 = \left(I - W^2\right)\left(I - W^2\right)^T$ respectively, as explained above for the general LLE case. This would be equivalent to minimizing

$$\operatorname{tr}\left(Y^1 - W^1Y^1\right)\left(Y^1 - W^1Y^1\right)^T + \operatorname{tr}\left(Y^2 - W^2Y^2\right)\left(Y^2 - W^2Y^2\right)^T . \tag{8}$$

In constrained LLE, the constraint that the parts of the two data sets which are in correspondence should have identical low-dimensional representation is used. Representing Y^1 and Y^2 as $Y^1 = \left[Y_c^1 Y_s^1\right]$ and $Y^2 = \left[Y_c^2 Y_s^2\right]$, the constraint is that $Y_c^1 = Y_c^2$. If M^1 and M^2 above are partitioned as

$$M^1 = \begin{bmatrix} M_{cc}^1 & M_{cs}^1 \\ M_{sc}^1 & M_{ss}^1 \end{bmatrix}, M^2 = \begin{bmatrix} M_{cc}^2 & M_{cs}^2 \\ M_{sc}^2 & M_{ss}^2 \end{bmatrix} \tag{9}$$

the cost in Equation (8) can be efficiently minimized under the constraint $Y_c^1 = Y_c^2$ by the eigenvectors of

$$M' = \begin{bmatrix} M_{cc}^1 + M_{cc}^2 & M_{cs}^1 & M_{cs}^2 \\ M_{sc}^1 & M_{ss}^1 & \mathbf{0} \\ M_{sc}^2 & \mathbf{0} & M_{ss}^2 \end{bmatrix} \tag{10}$$

corresponding to its smallest eigenvalues. The upper part of the eigenvectors would contain the low-dimensional embedding coordinates of the data from both sets which are in correspondence, while the bottom parts would correspond to the remaining data (the parts not in correspondence) from each set.

While the task of constrained LLE as introduced in [10] is to recover the missing high-dimensional data in (7), here we modify the algorithm to be able to perform local regression. The resulting method we call Locally Linearly Embedded Regression (LLER), and its mechanism is illustrated in Fig. 1.

Assume that we have a data set X and a corresponding parameter set P, among which some parameters may be unknown and the task is to estimate them. For example, in our case the data may contain the images of a certain object under different pose, and the parameters may represent the pose. For n_c of the images the corresponding pose is known, i.e. the correspondence $x_i \leftrightarrow p_i \ \ (i = 1, \cdots, n_c)$ is given. The task is to find the pose p_T of a test sample x_T. As before, we use the following block matrix format to represent both the available and missing (yet unknown) data.

$$X = \left[X_c^{\text{all}} \middle| X_s^{\text{all}}\right] = \left[\begin{array}{c|cc} X_c & x_T & X_s \\ P_c & p_T & P_s \end{array}\right] \tag{11}$$

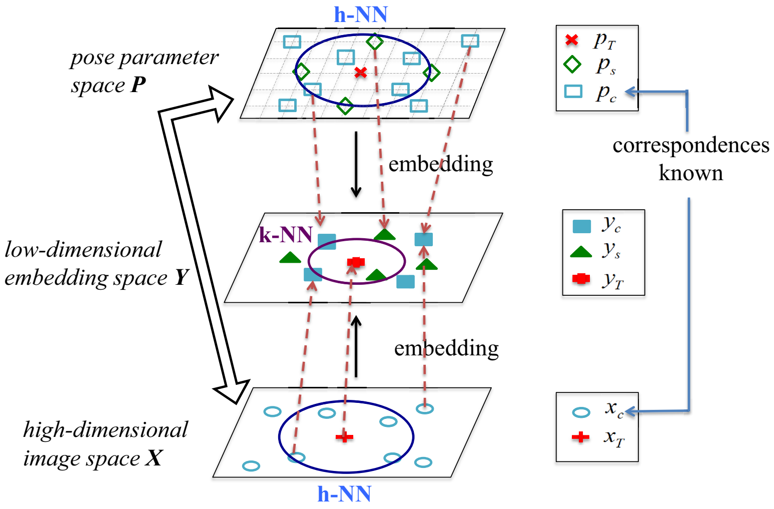


Fig. 1. Overview of the Locally Linearly Embedded Regression (LLER) algorithm

Here, X_c (object images) and P_c (corresponding pose parameters) are the training set, where the correspondence between image and pose is known, and x_T is the test image whose pose p_T needs to be found. P_s are pose parameters for which data (object images) are not available, i.e. X_s is missing data. Note that P_s can be easily generated on a grid of pose values (as shown in Fig. 1), sampled with a uniform step σ_i along each dimension i of pose space. The grid can take values in the same (or a little bit wider) range as the range in which the pose for the training data changes. The sampling step σ_i determines the density of the grid and generally should be taken to be smaller than the density of the available training parameters. For example, if the available training pose parameters for pan are sampled at 5 deg, σ_{pan} can be chosen to be 2.5 or 1.25, etc. However, taking too dense a grid P_s would naturally incur higher computational cost.

Separate (not constrained) low-dimensional representations of the data $Y^{(X)}$ and the parameters $Y^{(P)}$ can be obtained by diagonalizing respectively $M^{(X)} = \left(I - W^X\right)\left(I - W^X\right)^T$ and $M^{(P)} = \left(I - W^P\right)\left(I - W^P\right)^T$. However, by representing $M^{(X)}$ and $M^{(P)}$ as

$$M^{(X)} = \begin{bmatrix} M_{cc}^{(X)} & M_{cT}^{(X)} \\ M_{Tc}^{(X)} & M_{TT}^{(X)} \end{bmatrix}, M^{(P)} = \begin{bmatrix} M_{cc}^{(P)} & M_{cs}^{(P)} \\ M_{sc}^{(P)} & M_{ss}^{(P)} \end{bmatrix} \tag{12}$$

we can obtain a constrained embedding of the data and the parameters, represented by the eigenvectors corresponding to the smallest eigenvalues of the following matrix

$$M' = \begin{bmatrix} M_{cc}^{(X)} + M_{cc}^{(P)} & M_{cT}^{(X)} & M_{cs}^{(P)} \\ M_{Tc}^{(X)} & M_{TT}^{(X)} & \mathbf{0} \\ M_{sc}^{(P)} & \mathbf{0} & M_{ss}^{(P)}. \end{bmatrix} \tag{13}$$

In the resulting low-dimensional representation in the embedding space $\boldsymbol{Y}$, all training data x_c and parameters p_c which are in correspondence will map to the same point y_c, as indicated in Fig. 1 by the blue filled rectangles. The test data x_T which has no corresponding point in parameter space will map to y_T in $\boldsymbol{Y}$, and the parameters p_s which have no corresponding points in image space will map to the filled green triangles y_s. Then the LLER algorithm estimates the pose p_T of the test sample x_T through the following 3 steps[1].

The Locally Linearly Embedded Regression (LLER) Algorithm

step 1 Use constrained LLE with h-nearest neighbors for the weights $W^{(X)}$ and $W^{(P)}$ to find the constrained embeddings in $\boldsymbol{Y}$ for X_c, x_T and P_s, which are respectively Y_c, y_T and Y_s.

step 2 Find the k-nearest neighbors of y_T in $\boldsymbol{Y}$ and store them in a matrix Y_k. Store their corresponding pose parameters from pose space $\boldsymbol{P}$ into a matrix P_k (Note that some of the neighbors might be embeddings of images with known pose, while others might be embeddings of pose parameters without known corresponding images).

step 3 Find the linear map A between the embeddings Y_k and their corresponding pose parameters P_k using Equation (14) below, and use it to find the pose p_T of the test image x_T from Equation (15).

$$A = Y_k^+ P_k \tag{14}$$

$$p_T = y_T A = y_T Y_k^+ P_k \tag{15}$$

In Equations (14) and (15), Y_k^+ is the Moore-Penrose generalized inverse matrix [15], obtained by the singular value decomposition (SVD) of Y_k. Note that in the algorithm for LLER, in step 1 constrained LLE is performed on all available training data using h-nearest neighbors to determine the weights of the locally linear reconstructions, while in step 2 the k-nearest neighbors of the embedding of the test image in embedding space $\boldsymbol{Y}$ are used to determine the local neighborhood from which to obtain the estimation of the pose parameters. The parameters h and k need not take the same value (h would depend on the local structure of the image manifold, while k would depend also on the grid sampling step for P_s), and for that reason we have used different letters for them. Also, if the whole training data set is too large, performing cLLE on all data might be computationally expensive and actually would not be necessary as the following steps use only local information. Therefore, for huge training data sets only a subset of the whole training set which is in the larger neighborhood of the test sample (say not more than a few hundred samples) might be used to reduce computational cost.

[1] Matlab code for the LLER algorithm is available by writing to the corresponding author.

4 Experiments

In this section we evaluate LLER in comparison with several other global and local pose estimation methods using the Object Pose Estimation Database (OPED) [21], which seems to be the most accurate (with sufficient accuracy for robot grasping) publicly available data set [22]. The data set contains 703 different views for each of 16 different objects, sampled at 5 deg angle increments along two rotational axes (for pan between 0 and 180 deg, and tilt between 0 and 90 deg).

We perform two experiments, in each of which we randomly select respectively 100 or 350 images from each object to be used as a test sample pool, while the remaining images are used for training. This data splitting procedure is repeated randomly 5 times. For each test image the pose is estimated, compared to the available ground-truth pose values and the mean, standard deviation (std) and median of the estimation error is calculated. To facilitate comparison, the pan and tilt angles are represented as a vector on the unit sphere, so that in this way the absolute angle errors can be represented by a single value.

We compared LLRE to 3 global learning methods: nearest neighbor (NN) which simply selects the pose of the nearest neighbor to the query image, Linear Regression and Support Vector Regression (SVR) with RBF kernel. The local methods compared were LPR and LAR described in section 2. For all methods the relevant parameters were tuned to obtain best performance. For SVR, first reducing the feature dimensionality by PCA resulted in better performance, and this was used as a pre-processing step. For LLER, the parameter grid sampling step was chosen to be $\sigma_{pan} = \sigma_{tilt} = 2.5$ deg, the k and h for the k and h-nearest neighor selection were set to $h = 15$ and $k = 5$, and the dimension of the embeding space obtained from cLLE was $d = 10$.

The results obtained for 100 test images are shown in Table 1, and for 350 test images in Table 2. The results show that for most of the objects LLER achieves the smallest estimation error, in comparison with both the global learning methods and the other local learning methods. Also for LLER the error is low for all objects in the database, while SVR and LAR although performing well on some objects, for other objects (like bay, white car, house, socketin for SVR and bay, cap, ipipe for LAR) have quite a big estimation error.

5 Conclusion

In this paper we have proposed a novel local learning algorithm, Locally Linearly Embedded Regression, for appearance-based pose estimation. Our method embeds simultaneously the training images, the query image and a grid of pose parameters into an intermediate low-dimensional space, from which a linear map is learned between the points in the local neighborhood of the query representation and their corresponding pose parameters. which is used to directly recover the pose of the query image. The proposed method has been evaluated in a pose estimation task on a database of 16 different objects, on which it showed stable peformance across different objects and consistently outperfomed both other

Table 1. Experimental results (estimation error) for the case of 100 test samples per object; (top) global methods, (bottom) local methods

	NN			Linear Regression			SVR		
	Mean	STD	Median	Mean	STD	Median	Mean	STD	Median
adapter	3.48	1.87	5.00	4.25	3.75	3.24	0.47	0.43	0.34
bay	6.43	18.42	5.00	3.69	5.88	2.18	2.48	5.74	0.97
cablebox	3.44	1.80	5.00	1.62	1.16	1.38	0.36	0.37	0.27
cap	3.41	1.79	3.83	1.95	2.47	1.13	0.61	0.79	0.38
whitecar	3.47	1.84	5.00	3.81	3.30	2.92	1.82	2.55	0.97
clamp	3.84	2.09	5.00	3.21	2.89	2.45	1.25	1.29	0.81
fuse	3.67	2.00	5.00	2.09	1.83	1.62	1.04	1.60	0.58
goldcar	3.44	1.77	4.10	3.50	3.60	2.41	0.74	1.20	0.44
house	4.15	1.90	5.00	2.78	2.32	2.25	1.46	2.52	0.55
ipipe	4.56	6.34	4.53	1.68	1.89	1.28	1.02	1.62	0.60
redcar	3.48	1.80	4.53	4.31	3.70	3.29	0.72	0.59	0.56
socketin	3.74	1.89	5.00	3.90	4.34	2.60	1.27	1.39	0.86
socketout	3.78	2.01	5.00	2.70	3.63	1.57	0.72	0.93	0.45
tpipe	4.61	1.01	5.00	3.84	3.57	2.89	0.48	0.85	0.26
trap	3.75	1.93	5.00	1.72	1.77	1.18	0.67	1.02	0.36
wood	3.80	1.63	4.70	2.78	3.60	1.57	0.31	0.43	0.21
average	**3.94**	**3.13**	**4.79**	**2.99**	**3.11**	**2.12**	**0.96**	**1.46**	**0.54**

	LPR			LAR			LLER		
	Mean	STD	Median	Mean	STD	Median	Mean	STD	Median
adapter	1.18	1.35	0.71	0.70	0.70	0.48	0.47	0.54	0.26
bay	1.77	8.53	1.02	1.34	11.07	0.44	0.65	0.76	0.38
cablebox	0.81	1.03	0.49	0.56	0.60	0.40	0.39	0.46	0.24
cap	3.95	7.06	0.83	1.25	2.62	0.48	0.90	1.11	0.55
whitecar	0.78	0.88	0.50	0.61	0.64	0.45	0.42	0.60	0.24
clamp	1.68	1.98	1.04	1.71	2.54	0.80	0.76	1.04	0.42
fuse	1.38	1.73	0.71	0.54	0.63	0.36	0.41	0.57	0.19
goldcar	0.71	0.69	0.50	0.47	0.47	0.33	0.41	0.46	0.26
house	1.88	1.43	1.47	0.61	0.52	0.45	0.48	0.49	0.32
ipipe	1.85	7.16	0.64	0.79	3.67	0.40	0.45	0.71	0.24
redcar	1.03	1.40	0.69	0.65	0.56	0.50	0.50	0.56	0.33
socketin	1.88	3.76	0.73	0.78	1.93	0.43	0.58	0.93	0.31
socketout	1.87	5.36	0.68	0.70	5.18	0.24	0.51	1.01	0.29
tpipe	1.22	1.46	0.77	0.67	1.13	0.45	0.54	0.52	0.39
trap	1.73	2.89	0.84	0.47	0.52	0.31	0.55	0.93	0.25
wood	0.51	0.55	0.35	0.60	0.65	0.42	0.48	0.84	0.20
average	**1.52**	**2.95**	**0.75**	**0.78**	**2.09**	**0.43**	**0.53**	**0.72**	**0.30**

local learning algorithms and global regression algorithms like linear regression and non-linear SVR.

Also, the proposed LLER algorithm can be considered as a new type of general regression algorithm with a novel strategy — embedding together the data and a densely sampled grid of parameters (some of which do not correspond to any available data) in low-dimensional space, before obtaining a linear map between them, — and as such it might find applications apart from pose estimation in other problems where continuous-valued parameters need to be estimated from high-dimensional data like images.

Table 2. Experimental results (estimation error) for the case of 350 test samples per object; (top) global methods, (bottom) local methods

	NN			Linear Regression			SVR		
	Mean	STD	Median	Mean	STD	Median	Mean	STD	Median
adapter	3.93	2.25	5.00	5.43	4.86	4.00	1.18	1.69	0.65
bay	6.14	14.34	5.00	5.62	9.45	3.19	5.85	13.57	2.59
cablebox	3.60	1.81	4.98	2.30	1.80	1.82	0.80	1.34	0.42
cap	3.77	2.81	4.70	3.21	3.68	1.98	1.63	2.81	0.73
whitecar	3.72	1.89	5.00	5.08	4.61	3.70	4.58	6.31	2.39
clamp	4.35	2.44	5.00	4.48	4.17	3.32	2.25	2.93	1.35
fuse	4.23	2.72	5.00	3.07	2.96	2.15	3.49	5.24	1.51
goldcar	3.88	2.14	5.00	4.50	4.48	3.12	1.83	3.60	0.68
house	4.62	2.43	5.00	4.08	3.80	3.12	2.75	3.92	1.16
ipipe	5.44	9.55	4.92	2.58	3.01	1.82	2.82	7.19	0.85
redcar	3.85	1.84	5.00	5.03	4.01	3.89	1.51	1.77	0.95
socketin	4.06	2.75	5.00	4.77	5.38	3.12	1.98	4.18	0.77
socketout	4.40	4.40	5.00	3.49	5.16	2.05	1.55	3.88	0.54
tpipe	4.53	1.58	5.00	4.36	3.75	3.40	1.01	1.80	0.52
trap	4.09	2.35	5.00	2.56	2.61	1.85	1.79	3.72	0.72
wood	4.07	1.66	4.92	3.19	3.80	2.03	0.68	1.50	0.35
average	**4.29**	**3.56**	**4.97**	**3.98**	**4.22**	**2.78**	**2.23**	**4.09**	**1.01**

	LPR			LAR			LLER		
	Mean	STD	Median	Mean	STD	Median	Mean	STD	Median
adapter	1.72	1.58	1.19	1.33	1.58	0.85	0.88	1.04	0.56
bay	4.01	14.04	1.68	2.17	11.30	0.85	1.84	6.70	0.83
cablebox	1.43	1.36	1.06	1.18	1.51	0.76	0.94	1.31	0.57
cap	6.04	8.02	2.71	2.64	3.58	1.37	2.15	3.14	1.35
whitecar	1.68	5.05	0.96	1.26	1.53	0.81	1.53	2.14	0.98
clamp	2.99	3.47	1.88	2.77	4.13	1.27	1.74	2.85	0.88
fuse	2.30	2.46	1.41	1.13	1.33	0.68	0.97	1.38	0.54
goldcar	1.68	4.16	0.91	1.25	2.96	0.64	1.03	1.38	0.62
house	3.03	2.45	2.46	1.23	1.17	0.89	0.98	1.04	0.66
ipipe	3.55	13.58	1.06	2.20	9.25	0.79	1.21	3.72	0.63
redcar	1.78	2.11	1.22	1.22	1.46	0.85	1.14	1.41	0.71
socketin	3.25	12.40	1.14	1.99	7.98	0.67	1.26	1.91	0.72
socketout	3.20	12.65	1.10	1.61	7.53	0.44	1.44	5.19	0.60
tpipe	1.97	2.08	1.33	1.45	1.76	0.91	1.18	1.37	0.77
trap	2.20	3.35	1.32	1.09	1.60	0.61	1.10	1.48	0.59
wood	1.24	6.30	0.66	1.21	4.86	0.62	1.02	2.56	0.48
average	**2.63**	**5.94**	**1.38**	**1.61**	**3.97**	**0.81**	**1.28**	**2.41**	**0.72**

References

1. Grimson, E.: Object Recognition by Computer: The Role of Geometric Constraints. MIT Press (1990)
2. Lamdan, Y., Wolfson, H.J.: Geometric Hashing: A General and Efficient Model-Based Recognition Scheme. In: Proc. Int. Conf. on Computer Vision, Tampa, FL (1988)
3. Olson, C.F.: Efficient Pose Clustering Using a Randomized Algorithm. Int. Journal of Computer Vision 23(2), 131–147 (1997)
4. Lepetit, V., Fua, P.: Monocular Model-Based 3D Tracking of Rigid Objects: A Survey. Foundations and Trends in Computer Graphics and Vision 1(1), 1–89 (2005)
5. Murase, H., Nayar, S.K.: Visual Learning and Recognition of 3-D Objects from Appearance. Int. Journal of Computer Vision 14(1), 5–24 (1995)

6. Okatani, T., Deguchi, K.: Yet another appearance-based method for pose estimation based on a linear model. In: IAPR Workshop on Machine Vision Applications, pp. 258–261 (2000)
7. Melzer, T., Reiter, M., Bischof, H.: Appearance models based on kernel canonical correlation analysis. Pattern Recognition, Special Issue on Kernel and Subspace Methods for Computer Vision, 1961–1971 (2003)
8. Ando, S., Kusachi, Y., Suzuki, A., Arakawa, K.: Appearance based pose estimation of 3D object using support vector regression. In: International Conference on Image Processing, vol. 1, pp. 341–344 (2005)
9. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embeding. Science 290, 2323–2326 (2000)
10. Ham, J.H., Lee, D.D., Saul, L.K.: Learning High Dimensional Correspondences from Low dimensional Manifolds. In: The 20th International Conference on Machine Learning (2003)
11. Raytchev, B., Terakado, K., Tamaki, T., Kaneda, K.: Pose Estimation by Local Procrustes Regression. In: Proc. 18th IEEE International Conference on Image Processing, ICIP 2011, pp. 3666–3669 (2011)
12. Torgeson, W.: Multidimensional Scaling: I. Theory and method. Psychometrika 17, 401–419 (1952)
13. Cox, T., Cox, M.: Multidimensional Scaling, 2nd edn. Chapman & Hall/CRC (2000)
14. Mardia, K., Kent, J., Bibby, J.: Multivariate Analysis. Academic Press (1979)
15. Ben-Israel, A., Greville, T.N.E.: Generalized Inverses: Theory and Applications. Wiley, New York (1974)
16. Jolliffe, I.: Principal Component Analysis. Springer (1986)
17. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
18. Belkin, M., Niyogi, P.: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: Adv. NIPS 15, Vancouver, Canada (2001)
19. Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q.: Graph Embedding and Extensions: A General Framework for Dimensionality Rediction. IEEE Trans. PAMI 29(1), 40–51 (2007)
20. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
21. Viksten, F., Forssen, P., Johansson, B., Moe, A.: Comparison of Local Image Descriptors for Full 6 Degree-of-Freedom Pose Estimation. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1139–1146 (2009)
22. Object Pose Estimation Database, `http://www.cvl.isy.liu.se/research/objrec/posedb/`

Single-Channel Speech Dereverberation Based on Non-negative Blind Deconvolution and Prior Imposition on Speech and Filter

Il-Young Jeong, Biho Kim, and Hyung-Min Park

Department of Electronic Engineering, Sogang University, Seoul, Republic of Korea
{finejuly,biho,hpark}@sogang.ac.kr

Abstract. In this letter, we describe a single-channel speech dereverberation method in the short-time Fourier transform domain by using non-negative blind deconvolution. Robust decomposition of the magnitude spectra of the reverberated speech into its clean speech and a reverberation filter can be achieved by imposing a sparse and frequency-dependent prior model on the speech and an exponentially decaying envelope on the filter. Subsequently improved dereverberated speech is estimated without crude speech prior imposition for the fixed reverberation filter. The effectiveness of the algorithm was demonstrated with experimental results on speech reverberated by room impulse responses.

Keywords: Speech dereverberation, non-negative blind deconvolution, prior imposition.

1 Introduction

In real-world situations, audio signals such as speech are frequently corrupted by reverberation because they arrive at an observer through many paths, including reflections of walls. The farther the distance between a source and an observer or the more reverberant the environment is, the more severely deteriorated the quality of speech is.

Many methods for reducing the effect caused by reverberation have been developed. Nakatani et al. presented a short-time Fourier transform (STFT) domain approach to estimate dereverberated speech efficiently [1]. To make the problem simpler, the spectra of the reverberated signal are assumed to be the convolutive mixture of the spectra of clean speech and a filter in disregard of the effect of the phase components [2-4]. Especially, Kameoka et al. derived an efficient algorithm by applying a fast Fourier transform (FFT) to non-negative blind deconvolution and by imposing a clean speech prior which was modeled as a generalized Gaussian distribution [2]. Because the subband envelope of a clean signal is generally sparser than a reverberant one, the power spectrum of a reverberated speech signal could be decomposed into the estimated clean signal and a reverberation filter that causes reverberant tails.

Although the non-negative convolutive model for reverberated speech power spectra leads to a simple formulation, it may cause several drawbacks. In this letter, we

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 469–476, 2013.

propose a robust algorithm of single-channel speech dereverberation, while addressing these drawbacks. Since the power spectrum of reverberated speech is already quite sparse, this method employs the magnitude spectrum to magnify the difference in sparseness between the spectra of clean and reverberated speech. In addition, the assumption that the spectral components of a natural audio signal are independent across frequency is inappropriate, so we apply a frequency-dependent prior model which is more plausible to describe speech spectra. In addition, we assume that the filter has an exponentially decaying envelope, which is a typical shape of reverberation. Finally, a post-processing step is added to prevent excessive distortion caused by inaccurate speech prior for emphasizing sparseness.

2 Conventional Algorithm

Reverberated speech is generally modeled as convolution of its original clean speech and a reverberation filter. In the STFT domain, the reverberated speech at frequency bin k and frame t, $x_k[t]$, can be approximately represented by

$$x_k[t] \approx s_k[t] * h_k[t] = \sum_{\tau} s_k[\tau] h_k[t-\tau] , \tag{1}$$

where $s_k[t]$ and $h_k[t]$ denote the corresponding clean speech and reverberation filter at the same time-frequency segment, respectively, and $*$ denotes the convolution operator [2, 5]. Assuming that the values of $s_k[t]$ and $h_k[t]$ are uncorrelated, for a simple derivation, the power of $x_k[t]$ can be approximated as

$$\begin{aligned} X_k[t] &\approx |\sum_{\tau} s_k[\tau] h_k[t-\tau]|^2 \\ &\approx \sum_{\tau} S_k[\tau] H_k[t-\tau] \end{aligned} , \tag{2}$$

where $X_k[t]$, $S_k[t]$, and $H_k[t]$ denote $|x_k[t]|^2$, $|s_k[t]|^2$, and $|h_k[t]|^2$, respectively.

The power spectrum of the acquired speech signal in a reverberant environment, $Y_k[t]$, can be represented by the following generative model:

$$Y_k[t] = X_k[t] + \varepsilon_k[t] , \tag{3}$$

where $\varepsilon_k[t]$ denotes additive measurement noise or model error.

The main goal of the algorithm is to estimate $S \equiv \{S_k[1], \cdots, S_k[T]\}_{k=1}^{K}$ and $H \equiv \{H_k[1], \cdots, H_k[T]\}_{k=1}^{K}$ with observations $Y \equiv \{Y_k[1], \cdots, Y_k[T]\}_{k=1}^{K}$ by maximum-a-posteriori (MAP) estimation given by

$$\begin{aligned} \{S^*, H^*\} &= \arg\max_{S,H} P(S, H \mid Y) \\ &= \arg\max_{S,H} P(Y \mid S, H) P(S) P(H) \end{aligned} , \tag{4}$$

where S^* and H^* represent the optimal values of the clean speech and the reverberation filter. If there is no prior imposition on speech or filter, S^* and H^* will have

Y and the impulse response, respectively. With the assumption that the error in (3) is a zero-mean white Gaussian random variable, the likelihood term is expressed as

$$P(Y \mid S,H) = \prod_{k,t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_k[t]-X_k[t])^2}{2\sigma^2}\right) . \tag{5}$$

To model the sparseness of the clean speech, the generalized Gaussian prior is imposed to $P(S)$ as follows:

$$P(S) = \prod_{k,t} \frac{1}{2\Gamma(1+\frac{1}{p})b} \exp\left(-\frac{|S_k[t]|^p}{b^p}\right) , \tag{6}$$

where p should be set to a value in the interval of (0, 2) for $P(S)$ to be a super-Gaussian distribution. In contrast, $P(H)$ is assumed to be independent across frequency and follows a Dirichlet distribution whose parameters are all set to 1, so the distribution is the same as a uniform distribution. In addition, $\{H_k[1],\cdots,H_k[T]\}$ is normalized by $\sum_t H_k[t]=1$ to avoid indeterminacy in the scaling. With these priors, MAP estimation can be accomplished by solving the following optimization problem:

$$\begin{aligned} f(S,H) &= \sum_{k,t}(Y_k[t]-X_k[t])^2 + 2\lambda\sum_{k,t}|S_k[t]|^p \\ \text{subject to } &\sum_t H_k[t]=1, H_k[t]\ge 0, S_k[t]\ge 0 \quad , \end{aligned} \tag{7}$$

where λ determined by σ^2, b, and p is the weight parameter that represents the relative importance of the sparseness of S.

Since the non-negativity of S and H is premised, one can derive the solution by applying the non-negative matrix tensor deconvolution algorithm [6], which is an extended version of non-negative matrix factorization [7]. Therefore, the iterative rules for S and H are

$$\begin{aligned} S_k[\tau] &= S'_k[\tau]\frac{\sum_t H'_k[t-\tau]Y_k[t]}{\sum_t H'_k[t-\tau]X'_k[t] + \lambda p|S'_k[\tau]|^{p-1}} \\ &= S'_k[\tau]\frac{F^{-1}\{F\{H'_k[t]\}_u^* F\{Y_k[t]\}_u\}_\tau}{F^{-1}\{F\{S'_k[t]\}_u |F\{H'_k[t]\}|^2\}_\tau + \lambda p|S'_k[\tau]|^{p-1}} , \end{aligned} \tag{8}$$

$$\begin{aligned} H_k[\tau] &= H'_k[\tau]\frac{\sum_t S'_k[t-\tau]Y_k[t]}{\sum_t S'_k[t-\tau]X'_k[t]} \\ &= H'_k[\tau]\frac{F^{-1}\{F\{S'_k[t]\}_u^* F\{Y_k[t]\}_u\}_\tau}{F^{-1}\{|F\{S'_k[t]\}_u|^2 F\{H'_k[t]\}\}_\tau} , \end{aligned} \tag{9}$$

where $\{\cdot\}'$ denotes $\{\cdot\}$ at the previous iteration, and $F\{\cdot\}$ denotes the fast Fourier transform operator.

Although this method can be used to estimate clean speech, it may have some limitations. First, the algorithm operates on power spectra, such as $|x_k[t]|^2$ or $|s_k[t]|^2$, but $|x_k[t]|^2$ is too sparse to get the sparseness difference from $|s_k[t]|^2$. Therefore, $S_k[\tau]$

of (8) and $H_k[\tau]$ of (9) may converge into $Y_k[\tau]$ and the delta function, respectively. Furthermore, it is hard to increase the value of λ to extract the desired sparse clean speech because of the following side effects. Since spectral components are assumed to be independent across frequency, the estimated spectral powers of clean speech and the corresponding filters can be permuted with each other at a frequency bin. In particular, the spectral power of a reverberation filter at a frequency bin in the STFT domain is usually sparser than that of clean speech. Therefore, imposing a sparse prior on clean speech may lead to the problem that the spectral powers of the clean speech and the reverberation filter are estimated by (9) and (8), respectively, which is the opposite. In addition, the assumption of independent spectral components across frequency may cause an arbitrary frame-shift of the estimated clean speech at a frequency bin because $S_k[t]*H_k[t]=S_k[t+\tau]*H_k[t-\tau]$. This may result in the incorrect restoration of clean speech in the time domain. Moreover, the generalized Gaussian distribution is too inaccurate to describe the power spectrum of speech precisely, so estimated clean speech can be distorted by imposing the inaccurate prior.

3 Proposed Algorithm

In order to overcome these limitations, we present the following speech dereverberation algorithm. First, we derive the algorithm based on magnitude spectra to magnify the sparseness difference between clean and reverberated speech. In this section, $X_k[t]$, $S_k[t]$, and $H_k[t]$ denote $|x_k[t]|$, $|s_k[t]|$, and $|h_k[t]|$, respectively. The magnitude spectrum of reverberated speech, $X_k[t]$, is modeled as

$$\begin{aligned} X_k[t] &\approx |\sum_\tau s_k[\tau]h_k[t-\tau]| \\ &\approx \sum_\tau S_k[\tau]H_k[t-\tau] \end{aligned} . \tag{10}$$

In addition, the described method introduces a clean speech prior that models inherent dependencies across frequency, instead of using the conventional independent prior, which is given by

$$P(S) \propto \prod_t \exp(-\sqrt{\sum_k |\frac{S_k[t]}{\sigma_k}|^2}), \tag{11}$$

where σ_k represents the variance at the k-th frequency bin, which is set to 1 for convenience [8]. To avoid ambiguity causing an arbitrary shift of the estimated clean speech at a frequency bin, we also impose an exponentially decaying prior on the magnitude spectrum of the reverberation filter as follows:

$$P(H) \propto \prod_{k,t} \exp(-|\frac{H_k[t]}{\exp(-\varphi t)}|^2), \tag{12}$$

where φ is a constant related to the reverberation time (RT60). Practically, the value is not critical to the performance of the described method because the prior is imposed to avoid the arbitrary shift of the reverberation filter and the estimated clean speech by assuming an exponentially decaying shape. In this letter, we set it to 7, which approximately corresponds to a 1-s RT60. With these priors, the MAP estimation results in minimization of the cost function, with the same constraints in (7), expressed as

$$f(S,H)=\sum_{k,t}(Y_k[t]-X_k[t])^2+2\lambda\sum_t\sqrt{\sum_k|S_k[t]|^2}+2\mu\sum_{k,t}\left|\frac{H_k[t]}{\exp(-\varphi t)}\right|^2, \tag{13}$$

where $Y_k[t]$ denotes the magnitude spectrum of the observed signal.

It is noteworthy that the cost function contains a term derived from the prior of the reverberation filter in (12), whereas (7) does not contain the term because of the assumption of a uniform distribution. Similar to the derivation of non-negative blind deconvolution in the previous section, the magnitude spectra of clean speech and the reverberation filter can be iteratively estimated by

$$\begin{aligned} S_k[\tau] &= S_k'[\tau]\frac{\sum_t H_k'[t-\tau]Y_k[t]}{\sum_t H_k'[t-\tau]X_k'[t]+\lambda\frac{S_k'[\tau]}{\sqrt{\sum_k|S_k'[\tau]|^2}}} \\ &= S_k'[\tau]\frac{F^{-1}\{F\{H_k'[t]\}_u^*F\{Y_k[t]\}_u\}_\tau}{F^{-1}\{F\{S_k'[t]\}_u|F\{H_k'[t]\}_u|^2\}_\tau+\lambda\frac{S_k'[\tau]}{\sqrt{\sum_k|S_k'[\tau]|^2}}} \end{aligned}, \tag{14}$$

$$\begin{aligned} H_k[\tau] &= H_k'[\tau]\frac{\sum_t S_k'[t-\tau]Y_k[t]}{\sum_t S_k'[t-\tau]X_k'[t]+2\mu\frac{H_k'[t]}{\exp(-\varphi t)}} \\ &= H_k'[\tau]\frac{F^{-1}\{F\{S_k'[t]\}_u^*F\{Y_k[t]\}_u\}_\tau}{F^{-1}\{|F\{S_k'[t]\}_u|^2F\{H_k'[t]\}_u\}_\tau+2\mu\frac{H_k'[t]}{\exp(-\varphi t)}} \end{aligned}. \tag{15}$$

Although these update rules can decompose the magnitude spectrum of reverberated speech into those of clean speech and the reverberation filter, the estimated clean speech may have some error owing to the inaccurate clean speech prior in (11). To avoid the error, we re-estimate the clean speech with no prior imposition on clean speech and the fixed reverberation filter, after convergence of the update rules of (14) and (15). The update rule to re-estimate clean speech can be given by

$$\begin{aligned} S_k[\tau] &= S_k'[\tau]\frac{\sum_t H_k[t-\tau]Y_k[t]}{\sum_t H_k[t-\tau]X_k'[t]} \\ &= S_k'[\tau]\frac{F^{-1}\{F\{H_k[t]\}_u^*F\{Y_k[t]\}_u\}_\tau}{F^{-1}\{F\{S_k'[t]\}_u|F\{H_k[t]\}_u|^2\}_\tau} \end{aligned}. \tag{16}$$

With this rule, the magnitude spectrum of the estimated clean speech convolved with that of the reverberation filter becomes much closer to the observed magnitude spectrum of reverberated speech, and the spectral components of the clean speech that are excessively distorted are reasonably compensated.

4 Evaluation

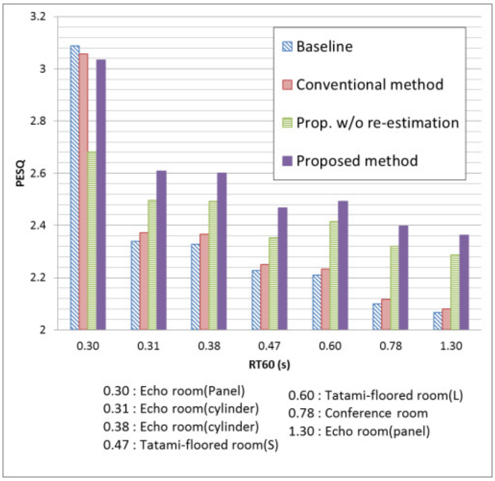
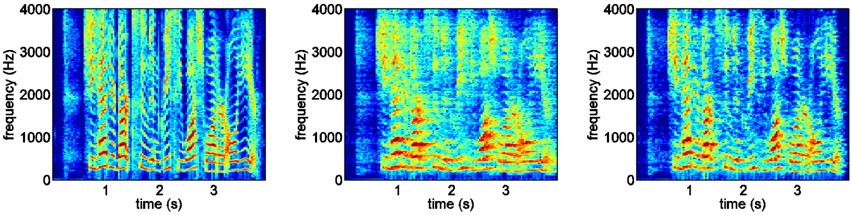


Fig. 1. Performance comparison of the PESQ scores between the original and estimated clean speech signals averaged over the TIMIT test dataset. For each reverberant environment, bars indicate the PESQ scores of the baseline, the conventional method in [2], the proposed method without re-estimation stage of (16), and the proposed method, respectively. Reverberation filters were simulated by RIRs selected from the RWCP Sound Scene Database in Real Acoustic Environments [9].



(a) Clean speech (b) Reverberated speech (c) Dereverberated speech

Fig. 2. Spectrograms of clean, reverberated, and dereverberated speech obtained using the described method. The clean speech signal was reverberated by the RIR corresponding to a 1.3-s RT60 in the RWCP Sound Scene Database in Real Acoustic Environments [9].

We evaluated the performance of the described method by using the TIMIT test dataset of 1680 sentences for the speech signal, and seven room impulse responses (RIRs) corresponding to different RT60s and environments selected from the RWCP Sound Scene Database in Real Acoustic Environments [9].

The described method was compared with the method in [2] in terms of the perceptual evaluation of speech quality (PESQ) between the original and estimated clean speech. Although the sampling rate of the original TIMIT data was 16 kHz, the data were down-sampled by a factor of 2 after anti-aliasing filtering to reduce the computational burden. The input signals were analyzed by a Hamming-windowed STFT with a 64-ms frame size and a 32-ms frame shift. Through extensive experiments, the optimal value of λ for the conventional method was set to E^{2-p}, where $E=\sum_{k,t}Y_k[t]\times 10^{-10}$, while λ for the described method was set to E^2, where $E=\sum_{k,t}Y_k[t]\times 10^{-5}$ (note that the conventional and described methods deal with the power and magnitude spectra, respectively; p was set to 1.2 as in [2]). For the described method, μ and φ were set to E^2, where $E=\sum_{k,t}Y_k[t]$, and 7 (as explained in Section III), respectively. H was initialized to be exponentially decaying over frame, and S to be the same as Y. Both the conventional and described methods run their source-filter deconvolutions for 20 iterations, and the described method runs the compensation rule of (16) for 5 extra iterations.

As shown in Fig. 1, the described algorithm shows better performance than the conventional method and the baseline for the RIRs at 0.31-s or greater RT60s. In the case of a 0.3-s RT60, the PESQ score of the baseline was already too high because the reverberation filter hardly deteriorated the speech quality. Even in this case, the PESQ score of the resulting speech from the described algorithm was slightly lower than, but comparable to, the baseline because the re-estimation process of the described algorithm efficiently reduced the side effect caused by aggressive parameter learning with the inaccurate clean speech prior. An example of the spectrograms of the original clean speech, reverberated speech, and dereverberated speech obtained using the described method is shown in Fig. 2. The figure demonstrates that the described method can remove reverberant components significantly. Some examples in the wave file format can be found at http://hompi.sogang.ac.kr/iip/research_derev.html, which confirm the effectiveness of the described method.

5 Conclusion

In this letter, we described a robust dereverberation method using non-negative deconvolution and prior imposition of speech and a reverberation filter. This method is based on the fact that the magnitude spectra of reverberated speech can be approximated as a convolutive mixture of clean speech and a reverberation filter. Furthermore, we imposed sparseness on the speech and an exponential decaying envelope on the filter for improving the efficiency and robustness of the algorithm. Finally, dereverberated speech was improved by post-processing without inaccurate prior imposition on the speech. Experimental results on speech reverberated by room impulse responses showed that reverberant components could be effectively removed by the described method.

Acknowledgements. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology of Korea (No.2010-0025703).

References

1. Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.-H.: Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation. In: IEEE ICASSP, pp. 85–88 (2008)
2. Kameoka, H., Nakatani, T., Yoshioka, T.: Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In: IEEE ICASSP, pp. 45–48 (2009)
3. Singh, R., Raj, B., Smaragdis, P.: Latent-variable decomposition based dereverberation of monaural and multi-channel signals. In: IEEE ICASSP, pp. 1914–1917 (2010)
4. Mysore, G.J., Smaragdis, P.: A convolutive spectral decomposition approach to the separation of feedback from target speech. In: IEEE Int. workshop on Machine Learning for Signal Processing (2011)
5. Krueger, A., Haeb-Umbach, R.: Model-based feature enhancement for reverberant speech recognition. IEEE Trans. Audio, Speech, Language Process. 18(7), 1692–1707 (2010)
6. Smaragdis, P.: Non-negative matrix factor deconvolution, extraction of multiple sound sources from monophonic inputs. In: Int. Workshop on ICA and BSS, pp. 494–499 (2004)
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Neural Information Processing Systems, pp. 556–562. MIT Press (2001)
8. Kim, T., Attias, H.T., Lee, S.-Y., Lee, T.-W.: Blind source separation exploiting higher-order frequency dependencies. IEEE Trans. Audio, Speech, Language Process. 15(1), 70–79 (2007)
9. Nakamura, S., Hiyane, K., Asano, F., Nishiura, T., Yamada, T.: Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In: Int. Conf. on Language Resources and Evaluation (2000)

Spectral Feature Extraction Using dNMF for Emotion Recognition in Vowel Sounds

Bo-Kyeong Kim and Soo-Young Lee

Dept. of Electrical Engineering,
Korea Advanced Institute of Science and Technology, Rep. of Korea
{kbghome,sylee}@kaist.ac.kr

Abstract. Recognizing emotional state from human voice is one of the important issues on speech signal processing. In this paper, we use the dNMF algorithm to find emotion-related spectral components in word speech. Each word consists of only vowels to remove language-dependent emotional factors. The dNMF algorithm with the additional Fisher criterion on the cost function of conventional NMF was designed to increase class-related discriminating power. Our experiment to recognize happiness, sadness, anger, and boredom in vowel sounds shows that more informative harmonic structures are computed by dNMF than NMF. Furthermore, dNMF features result in better recognition rates than NMF features for speaker-independent emotion recognition.

Keywords: discriminant NMF (dNMF), NMF, speech emotion recognition, speaker independent emotion recognition.

1 Introduction

Understanding human emotion from speech has attracted considerable attention in recent years, particularly in the field of human-robot interaction. In order to make robot to recognize which feelings and moods users have, the identification of efficient features which characterize different emotions is necessary [1].

Emotional state changes a vocal fold vibration in articulatory control and thus influences pitch harmonic structures in vowels. Pronouncing vowel sound /a/ in three emotional states has been reported to be beneficial for the classification of emotion [2]. Spectrum-based features using Non-negative Matrix Factorization (NMF) could be efficient tool for speech emotion recognition [3]. The discriminant NMF (dNMF) algorithm was also used to extract subtle emotional differences in speech [4]. However, dNMF in [4] was applied to the computed prosodic features requiring several computation and statistics.

In this paper, we employ dNMF directly on the time-frequency representation of raw speech signal. This will result in better performance than [3] as dNMF learns features to increase the discriminating power for classification. Moreover, we aim to discover emotional attributes on pitch harmonic structures of vowel sounds using dNMF. This paper is organized as follows: First, the principles of NMF and dNMF

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 477–484, 2013.

are briefly introduced. Next, we explain the word speech data and method used in the numerical experiment. Details in feature extraction and classification are also mentioned here. Finally, the analysis of feature vectors and recognition result for speaker-independent system are shown and discussed.

2 Background

2.1 NMF

The NMF algorithm is a popular feature extractor to factorize a non-negative $M \times N$ data matrix $\mathbf{X}$ into two non-negative matrices, a $M \times R$ basis matrix $\mathbf{W}$ and a $R \times N$ feature matrix $\mathbf{H}$ [5]. The columns in basis and feature matrix denote part-based building blocks and coefficients to explain how these blocks are linearly added to represent original data samples, respectively. The cost function using the square of Euclidean distance is

$$E_{NMF} = \frac{1}{2MN}\|\mathbf{X} - \mathbf{WH}\|^2 = \frac{1}{2MN}\sum_{m=1}^{M}\sum_{n=1}^{N}(X_{mn} - \sum_{r=1}^{R} W_{mr}H_{rn})^2 \tag{1}$$

Although NMF generally provides sparse representation of data, NMF learned by unsupervised learning does not have high discriminant power in classification.

2.2 discriminant NMF

The dNMF algorithm, proposed in [4], maximizes the Fisher linear discriminant of features while minimizing the NMF cost function simultaneously. The cost function E_{dNMF} is written as

$$E_{dNMF} = E_{NMF} - \lambda_D E_D \tag{2}$$

$$\text{where} \quad E_D = \frac{1}{2NR}\sum_{r=1}^{R}\sum_{k=1}^{K} N_k(\mu_{rk} - \mu_r)^2 \tag{3}$$

The terms for NMF representation error, discriminant power, and a relative weighting factor are denoted as E_{NMF}, E_D, and λ_D, respectively. The formulation of E_D is related to the between-class variance where μ_{rk} and μ_r are the mean coefficients of the k th class and of all samples for the r th feature.

The multiplicative update rule arising from gradient descent to minimize E_{dNMF} results in

$$W_{mr} \Leftarrow W_{mr}\frac{(\mathbf{XH}^T)_{mr}}{(\mathbf{WHH}^T)_{mr}} \tag{4}$$

$$H_{rn} \Leftarrow H_{rn}\frac{\left(\mathbf{W}^T\mathbf{X}+\lambda_D\frac{M}{R}\mathbf{HM}_c\right)_{rn}}{\left(\mathbf{W}^T\mathbf{WH}+\lambda_D\frac{M}{R}\mathbf{HM}_a\right)_{rn}} \tag{5}$$

where $(\cdot)_{mr}$ denotes the m th element of r th column of a matrix. Note that the update rule with $\lambda_D = 0$ is equivalent to NMF.

3 Method

3.1 Data Description

We develop a simulated emotional speech database that is collected from 14 professional voice actors (seven males and seven females). The database is comprised of 7 basic emotions (happiness, sadness, anger, boredom, disgust, fear, and surprise) as well as neutral one. For a given emotional state, 11 di- and 6 tri-syllabic words consisting of only vowels were recorded from each actor with 16 kHz sampling frequency. The use of vowel sounds is promising to investigate the language-independent emotional components related to pitch harmonics. Table 1 presents the phonetic symbols of words used in our database.

Table 1. Phonetic symbols used for word speech data

Words	Phonetic symbols
11 di-syllabic	/aa/, /ee/, /oo/, /uu/, /ii/, /au/, /ai/, /ei/, /ou/, /ua/, /ia/
6 tri-syllabic	/auu/, /aii/, /eii/, /ouu/, /uaa/, /iaa/

In this study, we use 136 word speeches for two-emotion recognition and 272 speeches for four-emotion recognition which are spoken by 4 selected male actors among the above database.

3.2 Preprocessing

We obtain the time-frequency representation $\boldsymbol{X_k}$ for k th word speech signal $x_k(n)$ of length L, by applying Short-Time Fourier Transform (STFT) as follows:

$$\psi\{x_k(n)\} = X_k(f,\tau) = \sum_{n=0}^{N-1} x_k(n) w(n-\tau) e^{-jnf} \tag{6}$$

where ψ and $w(n)$ denote the STFT operator and hamming window function of length N, respectively. Then the linear spectrogram incorporates the squared magnitude of $X_k(f,\tau)$ and is given as a matrix $\boldsymbol{X}_k = [|X_k(f,\tau)|^2] \in \mathbb{R}^{F \times T}$. Input data matrix for NMF and dNMF is constructed by collecting K vectorized spectrograms

$$\boldsymbol{X} = [vec(\boldsymbol{X}_1) \cdots vec(\boldsymbol{X}_K)] \in \mathbb{R}^{FT \times K} \tag{7}$$

where $vec(\boldsymbol{X}_k) \in \mathbb{R}^{FT \times 1}$ is a vector whose elements are taken columnwise from $\boldsymbol{X}_k$. Finally, we get the data matrices $\boldsymbol{X}_{train} \in \mathbb{R}^{FT \times K_{train}}$ and $\boldsymbol{X}_{test} \in \mathbb{R}^{FT \times K_{test}}$ obtained from K_{train} training samples and K_{test} testing samples, respectively.

In our experiment, the length of all samples is processed to be 750msec by zero-padding. For STFT, 20msec hamming window with frame shift interval of 10msec is used, and a windowed segment is Fourier-transformed by 1024-point FFT. We use the portion of each spectrogram only in the frequency range from 20Hz and 1000Hz leading the value of F to be 92 and normalize the sum of its power over 71 time-frames to be one.

3.3 Feature Extraction

We apply the dNMF algorithm in (4) and (5), to the training data matrix $\boldsymbol{X}_{train}$, leading to $\boldsymbol{X}_{train} = \boldsymbol{W}\boldsymbol{H}_{train}$ where $\boldsymbol{W} \in \mathbb{R}^{FT \times r}$ and $\boldsymbol{H}_{train} \in \mathbb{R}^{r \times K_{train}}$ are the basis and feature matrix. Here r is the reduced dimension containing temporal and spectral information of feature vectors. Notice that varying the value of λ_D in (5) controls the contribution of discriminant power on the cost function. Then the basis matrix $\boldsymbol{W}$ is used to infer associated features $\boldsymbol{H}_{test}$ by applying the algorithm (5) to testing data matrix $\boldsymbol{X}_{test}$ with $\lambda_D = 0$ since any class-related information is not provided in testing phase.

Figure 1 denotes the overall structure of our method. The representative speech characteristic in spectrogram such as pitch, harmonic structure, duration, and intensity can be clearly revealed in basis and feature matrix. Note that every vectorized spectrogram $\boldsymbol{x}$ (column of $\boldsymbol{X}_{train}$ or $\boldsymbol{X}_{test}$) is approximated by a linear combination of the columns of $\boldsymbol{W}$, weighted by the components of the corresponding feature vector $\boldsymbol{h}$ (column of $\boldsymbol{H}_{train}$ or $\boldsymbol{H}_{test}$).

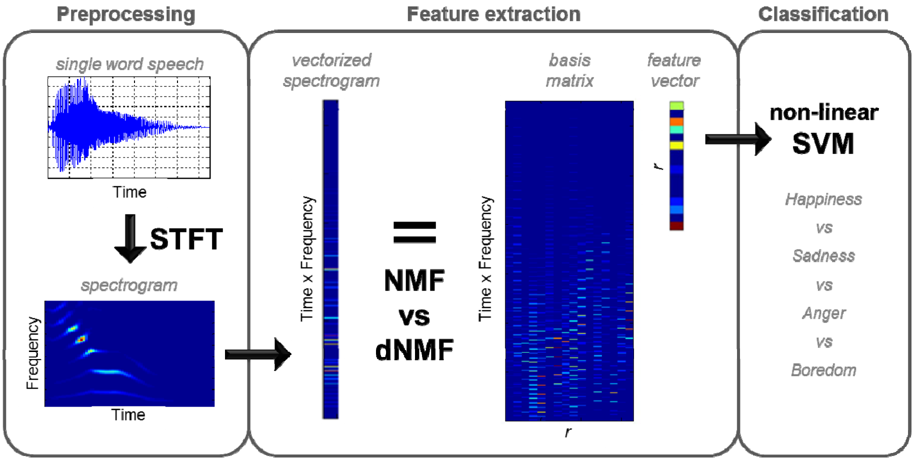


Fig. 1. Overall procedure of feature transformation for a word speech

3.4 Classification

The non-linear Support Vector Machine (SVM) with radial basis function (RBF) kernel is used as the classifier. For speaker-independent emotion recognition, we collect all word speeches from 4 speakers. The data samples are randomly divided into five sets for 5-fold cross validation. Optimal values for cost parameter of error and kernel width are found by a grid search [6].

It is widely known that classification between high-arousal and low-arousal emotion can be achieved at high accuracies, whereas classification among different emotions on the similar arousal level cannot [7]. First, we test our method to recognize 2 contrasting emotions (happiness and sadness) using 104 labeled and 32 unlabeled samples. After that, the 4 emotions (happiness, sadness, anger, and boredom) are classified using the SVM with one-against-one strategy. Notice that anger and happiness are both in high arousal level, and sadness and boredom are in low arousal level. The 4-emotion recognition utilizes 208 and 64 samples for training and testing, respectively.

4 Result and Discussion

4.1 Feature Extraction and Fisher Discriminant Score

The recorded signals of the same word in different emotional states show almost similar time-frequency representations, but subtle differences in pitch harmonics exist. The NMF and dNMF algorithms can basically extract the pitch harmonic structures in spectrogram of vowel sounds. The main issue is how to detect the differences coming from emotions and represent discriminant features for a better classification.

To prove the increase in discriminant power of dNMF, the Fisher discriminant score is plotted over 30000 epochs during the learning phase in Figure 2. Here the feature dimension of $r = 8$ for 4-emotion recognition is used. The Fisher linear discriminant values of feature coefficients are calculated for each dimension and summed for all dimensions at every 1000 learning epochs. The increasing tendency is clearly shown over learning epochs for dNMF with each value of $d\lambda$, whereas the tendency to remain unchanged is shown for NMF.

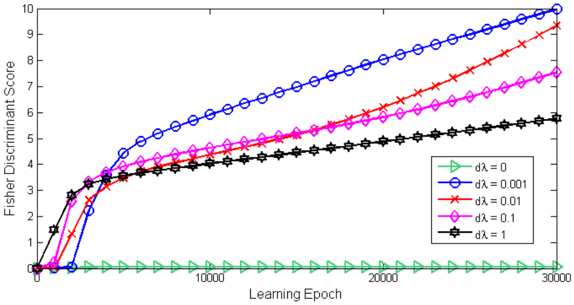


Fig. 2. Fisher discriminant scores as functions of learning epoch

Figure 3 illustrates the basis matrix and feature vector of word /ei/ acquired from the first speaker with $r = 8$ for 4-emotion recognition. Both NMF and dNMF present harmonic information in basis matrix and corresponding activation in components of feature vector. In this example, NMF features have the common activation corresponding to the 5th basis column for all 4 emotions. In contrast, dNMF features provide the distinguishable activation: the activation of 5th basis column for happiness, 3rd basis column for sadness, 7th basis column for anger, and 4th basis column for boredom. This discriminant characteristic of feature vector can lead to better recognition performance.

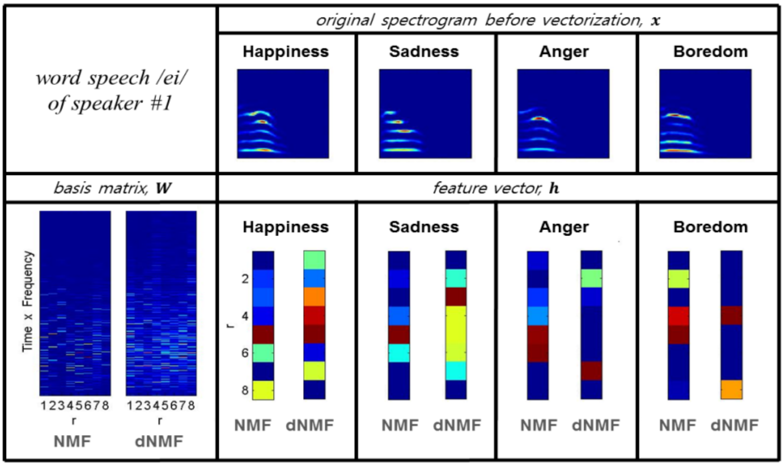


Fig. 3. An example of feature extraction using NMF and dNMF

4.2 Classification Result of Speaker-Independent Emotion Recognition

Figure 4 indicates the mean accuracy over 5-fold cross validation for 2-emotion recognition varying the value of $d\lambda$ and the feature dimension r. The dNMF algorithm provides the best classification result over all r's: 79.38% with $d\lambda = 0.01$ for $r = 2$, 80.00% with $d\lambda = 1$ for $r = 4$, 81.25% with $d\lambda = 0.001$ for $r = 10$, and 81.25% with $d\lambda = 0.01$ and 0.1 for $r = 50$. Especially, in the lowest dimension of $r = 2$, the average recognition rate of 79.38% is achieved by dNMF while much lower rate of 68.75% is achieved by NMF.

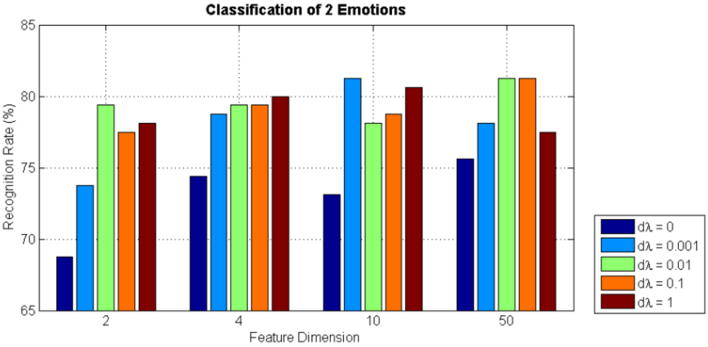


Fig. 4. Recognition result of 2 emotions (happiness and sadness) varying the value of $d\lambda$ for each feature dimension. Note that the case with $d\lambda = 0$ is equivalent to NMF.

The classification result of 4 emotions is shown in Figure 5. Similar to the results of 2-emotion recognition, the mean accuracies of 4-emotion recognition using dNMF outperform those using NMF over all r's. The maximum mean accuracy for each r is 54.38% with $d\lambda = 0.1$ for $r = 4$, 62.19% with $d\lambda = 0.001$ for $r = 8$, 58.75% with $d\lambda = 0.1$ for $r = 16$, 59.38% with $d\lambda = 0.01$ for $r = 64$, and 55.31% with $d\lambda = 0.001$ for $r = 128$.

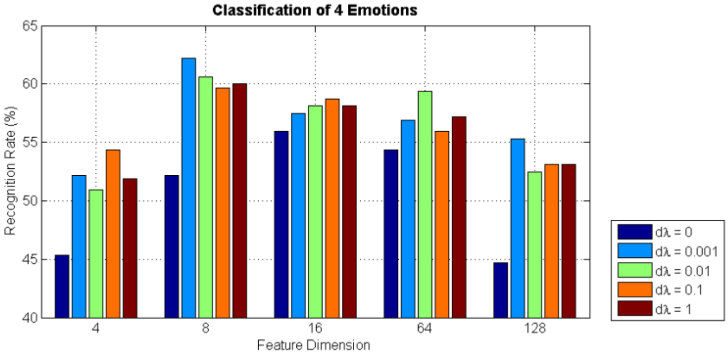


Fig. 5. Recognition result of 4 emotions (happiness, sadness, anger, and boredom) varying the value of $d\lambda$ for each feature dimension. Note that the case with $d\lambda = 0$ is equivalent to NMF.

Table 2 demonstrates the confusion matrix of 4-emotion recognition using NMF and dNMF with $d\lambda = 0.001$ for the feature dimension of $r = 8$. The classification rates of all emotions are higher when using dNMF than using NMF. Particularly, when classifying happiness and anger which are both high-arousal emotions, the lower error rates are achieved: the recognition of happiness as anger decreases from 22.50% of NMF to 16.25% of dNMF, and the recognition of anger as happiness drops from 27.50 % to 16.25%. It would show an effectiveness of dNMF for classification of emotions in the similar arousal level.

Table 2. Confusion matrix of 4-emotion recognition using NMF (left) and dNMF (right). The *H*, *S*, *A*, and *B* stand for happiness, sadness, anger, and boredom, respectively. Note that the bold-faced rates of 4 classes are used to calculate the average accuracy (Ave.).

Recognition result using NMF (%)				
	H	*S*	*A*	*B*
H	**53.75**	15.00	22.50	8.75
S	17.50	**45.00**	16.25	21.25
A	27.50	18.75	**41.25**	12.50
B	6.25	10.00	15.00	**68.75**
Ave.	**52.19**			

Recognition result using dNMF (%)				
	H	*S*	*A*	*B*
H	**62.50**	16.25	16.25	5.00
S	13.75	**52.50**	12.50	21.25
A	16.25	18.75	**57.50**	7.50
B	5.00	12.50	6.25	**76.25**
Ave.	**62.19**			

5 Conclusion

In this paper, dNMF is used to extract pitch harmonic structures of vowel sounds representing human emotions. For speaker-independent emotion recognition, dNMF successfully captures the spectral characteristics which are discriminant for emotional states but universal for all subjects. In addition, the recognition performances of dNMF outperform those of NMF by providing higher Fisher discriminant scores over features. Our framework shows the potential and promise of dNMF to extract efficient spectral features for speech emotion recognition.

In future work, applying dNMF to multi-emotional recognition and real-word data will be investigated. Since human speech is highly susceptible to the acoustic environment, data collected from the real-world situation should be used further. Also, facial expressions recorded in our constructed database will be utilized towards developing a multimodal recognition system.

Acknowledgement. This research was supported by the Brain Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning (grant 2013-035100).

References

1. Koolagudi, S.G., Rao, K.S.: Emotion Recognition from Speech: a review. Int. J. Speech Technol. 15, 99–117 (2012)
2. Tomas, B., Maletic, M., Raguz, Z.: Determination and Evaluation Pitch Harmonics Parameters with Emotions Classification. In: 15th International Conference on Software, Telecommunications and Computer Networks (2007)
3. Jeong, K.J., Song, J.Y., Jeong, H.: NMF Features for Speech Emotion Recognition. In: International Conference on Convergence and Hybrid Information Technology, pp. 368–374. ACM, New York (2009)
4. Lee, S.Y., Song, H.A., Amari, S.: A New discriminant NMF Algorithm and Its Application to the Extraction of Subtle Emotional Differences in Speech. Cogn. Neurodyn. 6, 525–535 (2012)
5. Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. In: Advances in Neural Information Processing Systems, pp. 556–562. MIT Press, Cambridge (2001)
6. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Department of Computer Science and Information Engineering, National Taiwan University, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
7. Ayadi, M.E., Kamel, M.S., Karray, F.: Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. Pattern Recogn. 44, 572–587 (2011)

Robust Visual Tracking Using Local Sparse Covariance Descriptor and Matching Pursuit

Bo Ma*, Hongwei Hu, Shiqi Liu, and Jianglong Chen

Beijing Lab of Intelligent Information Technology,
School of Computer Science and Technology, Beijing Institute of Technology
{bma000,huhongwei}@bit.edu.cn, shiqi6107@hotmail.com, ambjlon@163.com

Abstract. In this paper, we propose a visual tracking method based on local sparse covariance descriptor and matching pursuit. Covariance descriptor can model feature correlation of target templates effectively, and matching pursuit is employed to select the best target candidate which is reconstructed by target templates. The selection process is performed by solving a least square problem, and the candidate with the smallest projection error is taken as the tracking target. Experimental results on several video sequences demonstrate the good performance of proposed method compared with three existing tracking algorithms.

Keywords: Covariance descriptor, local sparse descriptor, visual tracking, matching pursuit.

1 Introduction

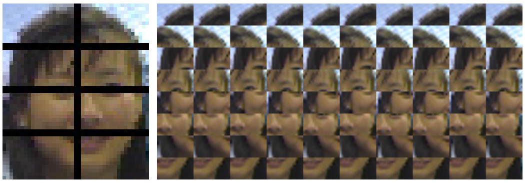
Visual tracking plays an important role in computer vision area. It is a challenging task to design a robust visual tracking algorithm, because target appearance often suffers from partial occlusion, background clutter, illumination changes, pose variation and shape deformation.

Lots of visual tracking algorithms [1,2] have been proposed in the last decade. Although these algorithms obtain good experimental results, how to model target appearance in real scenario is still a difficult problem. A discriminative and adaptive target appearance which can often be seen as the image features extracted from target is one of the most important parts in visual tracking. Different image features can be adopted to discriminate foreground object from background. Wang et al. [3] presented a tracking method in the perspective of mid-level vision with structural information captured in superpixels. Haar-like features are also widely used in many tracking algorithms [2, 4]. Target appearance can also be modeled by holistic templates [5] or local patches [6].

Recently, several tracking methods based on sparse representation with good performance proposed in [7–11] grabbed our attention. Mei et al. [8,12] regarded

* Supported by the National High-Tech Research & Development Program of China (2009AA01Z323) and the National Natural Science Foundation of China (Grant No. 61271374).

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 485–492, 2013.



(a) Candidate (b) Template set

Fig. 1. The candidate appearance patches in (a) are modeled by the template patch set in (b)

tracking as a sparse approximation problem. In their algorithm, each target candidate was sparsely represented in the space spanned by target templates and trivial templates. Zhong et al. [10] tracked object by a collaborative model including sparsity-based discriminative classifier and sparsity-based generative model. Jia et al. [9] argued that the target appearance could be represented by a structural local sparse model, and an alignment-pooling method was used in the sparse codes. Liu et al. [7] developed a tracking method based on local appearance model which located object by a sparse constraint regularized mean-shift algorithm.

In our method, region covariance descriptor extracted from a local image patch is taken as the representation. We assume that target is a linear combination of a small number of samples from a given training set during tracking process. Rather than solving a $l1$ convex optimization [13] to achieve sparsity, our method locates the target in newly arrived frame with a greedy manner by using matching pursuit [14–16]. The sparsity coefficient required in $l1$ tracker [12] is replaced by a maximal number of samples used in reconstructing the candidate target in order to reduce computational cost.

Our main contribution is concluded as follows: (1) We present a sparse covariance feature for template representation in visual tracking. (2) Matching pursuit is employed to select the closest approximation in target template set by solving a least square problem. (3) A template update scheme is proposed in our approach in order to handle the appearance changes.

2 Local Sparse Covariance Descriptor

2.1 Region Covariance Descriptor

Given an image I, φ is a mapping function that extracts a n-dimensional feature vector $\boldsymbol{z}_i \in \mathbb{R}^n$ for each pixel in I,

$$\varphi\left(I, x_i, y_i\right) = \boldsymbol{z}_i, \tag{1}$$

where (x_i,y_i) is the location of the i^{th} pixel. We represent the region R by a $n \times n$ covariance matrix C_R using these feature vectors $\{\boldsymbol{z}_i\}_{i=1}^{|R|}$,

$$C_R = \frac{1}{|R|-1} \sum_{i=1}^{|R|} (\boldsymbol{z}_i - \boldsymbol{\mu}_R)(\boldsymbol{z}_i - \boldsymbol{\mu}_R)^T, \tag{2}$$

where $|R|$ represents the number of pixels in region R and $\boldsymbol{\mu}_R = \frac{1}{|R|} \sum_{i=1}^{|R|} \boldsymbol{z}_i$.

In our experiments, feature vector $\boldsymbol{z}$ has the following form

$$\left[R, G, B, |I_x|, |I_y|, |I_{xx}|, |I_{yy}|, |I_{xy}|, \sqrt{I_x^2 + I_y^2}\right]^T, \tag{3}$$

where R, G, B are the three components of color channel. I_x, I_y and I_{xx}, I_{yy}, I_{xy} are the first and second order spatial derivatives of the image intensity with respect to x and y respectively.

Typically, covariance matrices don't lie in Euclidean space. Instead, they lie in a nonlinear Riemannian manifold [17,18]. By adopting Logarithm operator, the projected covariance matrices lie in an Euclidean space approximately. Therefore, we calculate the logarithmic matrix [19] for each covariance matrix. After that, the lower triangular elements of this resulting matrix are written in form of a column vector which can be seen as the representation of a region.

2.2 Local Sparse Representation

Sparse representation has been widely used in many research fields including visual tracking [7,9,12]. Given a template set of the tracking object, the target candidate can be represented as a linear combination of a few basis vectors in the set. Let $D = \{\boldsymbol{d}_j\}_{j=1}^{Q}$ as a dictionary consisting of Q m-dimensional vectors, and each $\boldsymbol{d}_j$ represents a template. Given a target candidate $\boldsymbol{t}$, it can be reconstructed by

$$\boldsymbol{t} \approx \sum_{j=1}^{Q} a_j \boldsymbol{d}_j + \boldsymbol{E}, \tag{4}$$

where $\boldsymbol{E}$ is the residual approximation error. The coefficient vector $\boldsymbol{a} = (a_1, \cdots, a_Q)^T$ which typically provides a sparse solution can be computed by optimizing the following problem,

$$\min_{\boldsymbol{a}} \left\| \boldsymbol{t} - \sum_{j=1}^{Q} a_j \boldsymbol{d}_j - \boldsymbol{E} \right\|_2^2 + \lambda \|\boldsymbol{a}\|_1, \tag{5}$$

where λ controls the balance between a sparser representation and a closer approximation.

In current frame, we can sample a set of candidates inside a region which is given by the tracking result in previous frame. For each candidate, we sample a set of small image patches $S\!=\!\{S_i|i\!=\!1\!:\!P\}$, as shown in Fig. 1(a), and the target is segmented into P patches. Similarly, for the corresponding template patch set, we have $A\!=\!\{A_i\}_{i=1}^{P}$, and $A_i\!=\!\{\boldsymbol{A}_{ij}|j\!=\!1\!:\!Q\}$ is shown in Fig. 1(b). Then, the objective function in (5) is changed as

$$\min_{x} \sum_{i=1}^{P} \left(\|S_i - x_{i1}\boldsymbol{A}_{i1} - x_{i2}\boldsymbol{A}_{i2} - \cdots - x_{iQ}\boldsymbol{A}_{iQ} - \boldsymbol{E}_i\|_2^2 + \lambda \|\boldsymbol{x}_i\|_1 \right), \tag{6}$$

where λ is a small constant and $x\!=\!\{\boldsymbol{x}_i\}_{i=1}^{P}$, $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{iQ})^T$ is the corresponding coefficient.

Different with the $l1$ tracker [12] who aims to alleviate the partial occlusion problem by incorporating the trivial templates at the price of high computation burden, in our algorithm, both target candidates and templates are segmented into patches, and each patch is represented by its covariance descriptor. With a local sparse representation, the candidate patch appearance is modeled by a sparse template patch set. Therefore, when the target is occluded partially, some occluded patches cannot match with templates correctly, but the rest patches may help get accurate localization of the object.

3 Tracking by Matching Pursuit

Many methods have been proposed to minimize the overall reconstruction error for sparse representation [7, 8, 12, 13]. Instead of solving a $l1$ optimization problem which usually requires high computational cost, we resort to matching pursuit [20] to speed up the tracking algorithm without sacrificing the tracking performance.

We rewrite (6) as

$$\min_{x,T} \sum_{i=1}^{P} \|S_i - x_{i1}\boldsymbol{T}_{i1} - x_{i2}\boldsymbol{T}_{i2} - \cdots - x_{iK}\boldsymbol{T}_{iK}\|_2^2, \tag{7}$$

where S_i is the descriptor of i^{th} patch of a candidate, $T\!=\!\{\boldsymbol{T}_{ij}\}_{j=1}^{K}$ the solution set which is selected from the i^{th} template patch set A_i by matching pursuit, and x_{ij} is the corresponding coefficient of $\boldsymbol{T}_{ij}$.

Now we consider choosing the $P \times K$ template patches one by one. Denote $\boldsymbol{A}_{ij} = (a_1, a_2, \cdots, a_m)^T \in \mathbb{R}^m$ and $S_i = (s_1, s_2, \cdots, s_m)^T \in \mathbb{R}^m$. For each $S_i\,(i = 1, \cdots, P)$ and $A\!=\!\{\boldsymbol{A}_{ij}\}_{i=1}^{P}\,(j = 1, \cdots, Q)$, we have

$$\min_{x_{ij}} \|S_i - x_{ij}\boldsymbol{A}_{ij}\|_2^2 \Longrightarrow \min_{x_{ij}} \sum_{z=1}^{m} (s_z - x_{ij}a_z)^2. \tag{8}$$

Denote $f(x_{ij}) = \sum\limits_{z=1}^{m} (s_z - x_{ij}a_z)^2$ and let $\frac{\partial f(x_{ij})}{\partial x_{ij}} = 0$. We have

$$x_{ij} = \frac{\sum\limits_{z=1}^{m} a_z s_z}{\sum\limits_{z=1}^{m} a_z^2} = \frac{\boldsymbol{A}_{ij}^T S_i}{\|\boldsymbol{A}_{ij}\|^2}. \tag{9}$$

With each candidate patch S_i, we can compute a x_{ij} for $\boldsymbol{A}_{ij}$. We substitute (9) to (8); then, the template patch which corresponds to the smallest function value is selected as the best matching, and we denote it by $\boldsymbol{T}_{i1}$. To select the second item, we let $S_i{=}S_i{-}x_{i1}\boldsymbol{T}_{i1}$, $A_i{=}A_i - \{\boldsymbol{T}_{i1}\}$, and repeat this process in order to select T_{i2}. Similarly, all the best matching $\boldsymbol{T}_{i1}, \ldots, \boldsymbol{T}_{iK}$ can be chosen for each candidate patch S_i. By repeating these processes for other patches, $P \times K$ template patches will be selected. Substituting all these to (7), the target candidate which gets the smallest function value will be chosen as the tracking result.

4 Template Update

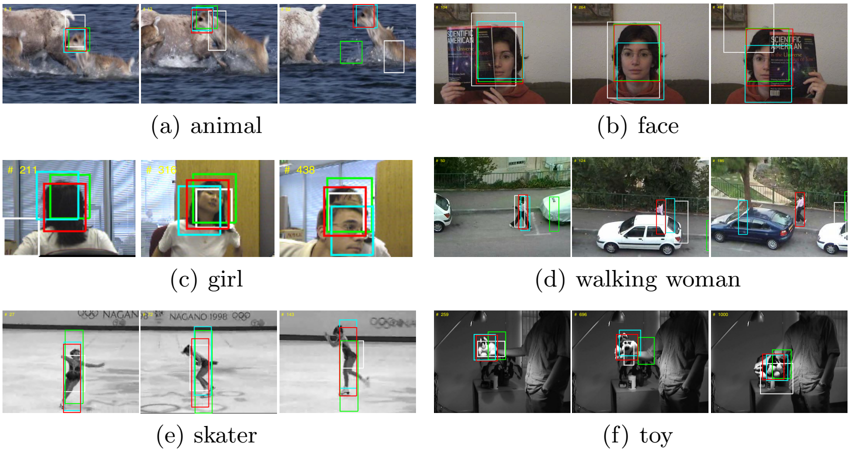
Target appearance will change after a period of time. If the method use a static template set, eventually it is not able to model object appearance accurately when it suffers from changes, such as illumination changes, partial occlusion, clutter or shape deformations. However, if the template is updated too frequently with new observations, errors are likely to accumulate and the tracker will drift away from the target. In our algorithm, we dynamically update the target template set A to tackle this problem. A mechanism should be adopted to decide on when and how to update the template to balance between capturing the appearance changes of the target and reducing the error accumulation.

Initially, the first target template is manually selected in the first frame and the rest target templates are created by perturbing it near the true position. The target template set A is then updated every E frames. Considering which template should be replaced when updating, we set a counter for each of them. The counters are set to zero with the initial template set. When we get the tracking result for each frame, add one to each corresponding counter of templates whose patches are selected by matching pursuits for the chosen target candidate. For doing update, the template in set A which has the minimum value in its corresponding counter will be replaced by the tracking result of current frame.

5 Experiments

Our proposed method is compared with three other algorithms including covariance tracker (COV) [21], multiple instances learning (MIL) [2], and visual tracking using l1 minimization($l1$) [12]. We set the template number Q to 10 and optimal matching number K to 3. For 24-bit color image sequences, the number of covariance feature dimension is $n = 9$ in (3). The patch number P is set to 8. While for 8-bit gray scale image sequences, the feature dimension n is set 7 and P is 16.

The first test sequence (animal) in Fig. 2(a) presents the tracking results where the target appears in background clutters. The tracking results with red, green, blue and white rectangles are for our proposed tracker, $l1$ tracker, MIL tracker and COV tracker, respectively. In Fig. 2(a) we can see that the color and



(a) animal (b) face

(c) girl (d) walking woman

(e) skater (f) toy

Fig. 2. Tracking results. Our tracker, $l1$ tracker, MIL tracker, COV tracker are represented by red, green, blue, and white rectangles respectively

textures of the foreground and background are very close to each other, so the target is hard to recognize. Only our method and MIL tracker can get correct results, while $l1$ tracker and COV tracker both lose the target.

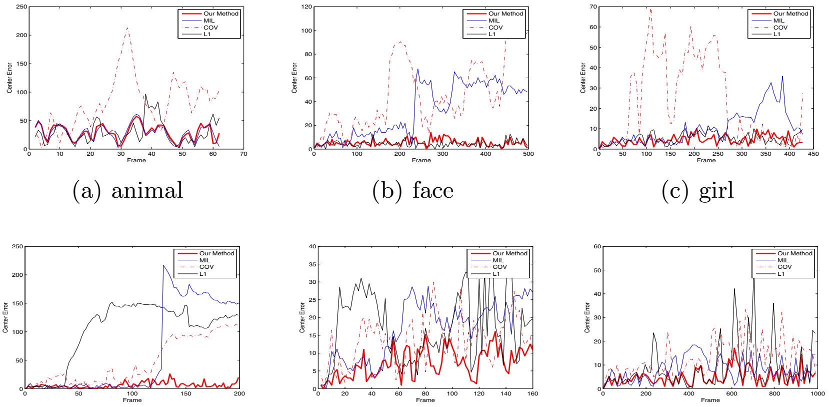
The second and the fourth test sequences (face and walking woman) show the target being occluded heavily. In Fig. 2(b) we see that the face of the woman is occluded by a book severely, and only our tracker and $l1$ tracker are capable of tracking the object all the time. In comparison, COV tracker loses the target in frame 204 and MIL tracker loses the target in frame 264. In Fig. 2(d), when the woman is passing the car, other trackers lose the target at frame 124 and then go out of range. Only our method obtains good tracking results throughout the whole sequence.

The third test sequence (girl) is under occlusion and large appearance variations. In Fig. 2(c) we can see that the girl is occluded by a man in frames 438, and our method gets the best tracking results in all four algorithms. Although $l1$ and MIL can track the target when it has appearance changes in frames 211, 292, 316, 328 and 390, they do not locate very well. Compared with our tracker, these two methods are less accurate when object suffers from viewpoint variations.

The fifth test sequence (skater) is under large pose variation and shape deformation. In Fig. 2(e) we can see that the athlete has many body movements like rotation, bent down, kick and stretch. All these movements make the shape of the skater keep changing and difficult to track. The results show that our method is more robust than COV, $l1$ and MIL trackers in appearance changes.

The last test sequence (toy) is under appearance and illumination changes. It can be observed that our method gets a more robust and accurate result than other trackers when the toy is rotated by a man. Some tracking result frames are given in Fig. 2(f).

At last, in Fig. 3, we present the relative position errors (in pixels) between the ground truth center and the tracking results of these four algorithms. It



(a) animal (b) face (c) girl

(d) walking woman (e) skater (f) toy

Fig. 3. Quantitative evaluation of the trackers in terms of position errors (in pixels)

Table 1. Average execution time per frame(sec)

	animal	face	girl	woman	skater	toy
MIL	0.2527	0.7839	0.3178	0.5530	0.3179	1.0838
COV	6.3243	7.1871	4.6742	6.5878	6.4937	6.5072
L1	3.2796	3.0077	3.3857	3.2096	3.1341	3.0082
OUR	2.7569	1.9080	1.6693	1.6001	2.6085	2.9914

shows our method produces almost the smallest tracking errors for all sequences. Table 1 shows the average execution time per frame obtained by four algorithms of the six sequences. In comparison, our tracker implements faster than L1 and COV trackers but slower than MIL tracker. All methods are implemented using MATLAB and performed on the same PC.

6 Conclusions

In this paper, we develop a robust tracking algorithm with a dynamic local sparse covariance dictionary. The target appearance is represented by covariance descriptor. Given the covariance training set of object of interest, our method aims at finding the optimally matched region that meanwhile satisfies the sparsity constraint. Local patches based representation helps alleviate the occlusion problem. Computational cost is reduced by using matching pursuit rather than solving a $l1$ convex optimization problem. A simple template set update scheme is adopted to remedy appearance change and drift problem. Promising experimental results have been reported by comparing with other up to date tracking algorithms.

References

1. Ross, D., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. IJCV 77, 125–141 (2008)
2. Babenko, B., Yang, M., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE CVPR, pp. 983–990 (2009)
3. Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: IEEE ICCV, pp. 1323–1330 (2011)
4. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
5. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR, vol. 1, pp. 798–805 (2006)
6. Hong, X., Chang, H., Shan, S., Zhong, B., Chen, X., Gao, W.: Sigma set based implicit online learning for object tracking. IEEE SPL 17, 807–810 (2010)
7. Liu, B., Huang, J., Yang, L., Kulikowsk, C.: Robust tracking using local sparse appearance model and k-selection. In: IEEE CVPR, pp. 1313–1320 (2011)
8. Mei, X., Ling, H., Wu, Y., Blasch, E., Bai, L.: Minimum error bounded efficient l 1 tracker with occlusion detection. In: IEEE CVPR, pp. 1257–1264 (2011)
9. Jia, X., Lu, H., Yang, M.: Visual tracking via adaptive structural local sparse appearance model. In: CVPR, pp. 1822–1829 (2012)
10. Zhong, W., Lu, H., Yang, M.: Robust object tracking via sparsity-based collaborative model. In: CVPR, pp. 1838–1845 (2012)
11. Zhang, S., Yao, H., Sun, X., Lu, X.: Sparse coding based visual tracking: review and experimental comparison. PR 46, 1772–1788 (2012)
12. Mei, X., Ling, H.: Robust visual tracking using l 1 minimization. In: IEEE CVPR, pp. 1436–1443 (2009)
13. Sivalingam, R., Boley, D., Morellas, V., Papanikolopoulos, N.: Tensor sparse coding for region covariances. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 722–735. Springer, Heidelberg (2010)
14. Bai, T., Li, Y.: Robust visual tracking with structured sparse representation appearance model. PR (2011)
15. Liu, B., Yang, L., Huang, J., Meer, P., Gong, L., Kulikowski, C.: Robust and fast collaborative tracking with two stage sparse optimization. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 624–637. Springer, Heidelberg (2010)
16. Li, H., Shen, C., Shi, Q.: Real-time visual tracking using compressive sensing. In: IEEE CVPR, pp. 1305–1312 (2011)
17. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. IJCV 66, 41–66 (2006)
18. Xueliang, Z., Bo, M.: Gaussian mixture model on tensor field for visual tracking. SPL 19, 733–736 (2012)
19. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine 56, 411–421 (2006)
20. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. SP 41, 3397–3415 (1993)
21. Porikli, F., Tuzel, O., Meer, P.: Covariance tracking using model update based on lie algebra. In: CVPR, vol. 1, pp. 728–735 (2006)

PCA-Based Appearance Template Learning for Contour Tracking

Bo Ma*, Hongwei Hu, Pei Li, and Yin Han

Beijing Lab of Intelligent Information Technology
School of Computer Science and Technology, Beijing Institute of Technology
{bma000,huhongwei,lipei00}@bit.edu.cn, ieyinhan@126.com

Abstract. A novel method is proposed in this paper to model changes of object appearance for object contour tracking. Principal component analysis is utilized to learn eigenvectors from a set of the object appearance in our work, and then the current object appearance can be reconstructed by a linear combination of the eigenvectors. To extract the object contour, we perform covariance matching under the variational level set framework. The proposed method is tested on several sequences under large variations, and demonstrates that it outperforms current methods without updating the appearance template.

Keywords: appearance template, PCA, contour tracking, level set, covariance matrix.

1 Introduction

Object contour tracking in video sequence is a very important part of computer vision. Contour-based methods aim to obtain the accurate contour of an object in each frame instead of the rough locations. Although contour tracking has been studied for many years, robust and accurate tracking of a deforming, non-rigid and fast moving object with appearance changes is still a challenging problem.

In recent years, the level set method is used to represent the object boundary, which ideally provides precise localization of the object [1–5]. In the level set method, object contour is represented by the zero level set of a high dimensional embedding function. In [1], Nikos Paragios et al. used a model-free approach which was robust to the presence of noise for tracking coping with important local deformations. The use of the level set method within such a framework leads to an implicit and parameter free approach that can cope with topological changes.

To model a target, one of the difficulties is the representation of the target region. In some literatures, the region is often described by known distributions or intensity histograms [6, 7]. Histogram-based descriptors integrate the statistical

* Supported by the National High-Tech Research & Development Program of China (2009AA01Z323) and the National Natural Science Foundation of China (Grant No. 61271374).

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 493–500, 2013.

information of an image region. As a result they are insensitive to spatial structure. Sevilla-Lara et al. [8] used distribution fields descriptor for tracking, which allowed smoothing objective function without destroying information about pixel values. This provides an effective way to overcome occlusions or small misalignments. Many researchers also use covariance matrix as region descriptor [9–12]. Tuzel et al. [9] proposed the covariance region descriptor to model the object appearance representation, which was capable of fusing correlated features inside an object region and invariant to uniform illumination, view and pose changes.

How to deal with the appearance variability of the target is another significant difficulty in visual contour tracking. Shaoting Zhang et al. [13] proposed a sparse shape composition model that adaptively approximated the input shape by a sparse linear combination of training shapes instead of explicitly learning shape priors. David A. Ross et al. [14] presented a tracking method that incrementally learned a low-dimensional subspace representation, adapting to appearance changes of target. In [10], covariance descriptor is adopted as appearance model. It simplifies a complex model update process on Riemannian manifold by computing weighted sample covariance that can be updated incrementally during object tracking process. Principal Component Analysis (PCA) has been of great interest in computer vision and pattern recognition [15, 16]. Li [15] constructed a subspace-based background model in which an online PCA was used to incrementally learn a background subspace representation.

In this paper, we propose a new method to model the changes of the object appearance and extract the accurate contour of object under the variational level set framework. Covariance matrix [9] is used as region-level feature descriptor in our method. From a set of training templates, we learn a set of eigenvectors using PCA. Then, the current object appearance can be represented by the linear combination of the eigenvectors. Therefore, the template of the object appearance will change in different frames. However, covariance matrices do not lie in Euclidean space. Therefore, logarithmic Euclidean distance [17] is adopted to measure the similarity between different covariance matrices. To extract the object contour, we minimize the similarity between the candidate region covariance matrix and the template meanwhile maximize the dissimilarity between the background region covariance matrix and the template.

The contributions of this paper are concluded as follows: (1)We propose a PCA-based target contour appearance representation. (2)A new energy functional is proposed to model the change of target contour. (3)A gradient descent algorithm is utilized to solve the energy functional.

2 Our Method

In this section, we will present the details of our new method. Firstly, we introduce the region descriptor covariance matrix which uses image second order statistics. Then, we learn a template which changes dynamically in different frames from a set of training data. Finally, we use this template to form the image energy and derive its gradient flow equation to evolve the contour of the object.

2.1 Region Descriptor

In our tracking framework, an object is represented by a covariance matrix of the image features inside the object region. Let I be an one dimensional intensity or three dimensional color image of size $W \times H$, and (x, y) denotes pixel coordinates. In our method for contour tracking, $f(x, y)$ is defined by pixel locations (x, y), image gray level or color and the norm of the first and second order derivatives of the intensities with respect to x and y,

$$f(x,y) = \left[x, y, I(x,y), I_x, I_y, I_{xx}, I_{yy}, I_{xy}, \sqrt{I_x^2 + I_y^2}\right]^T . \tag{1}$$

For a target region $R \subset \Omega$ (the image plane $\Omega = R \cup R^C$), it can be represented by a $d \times d$ (here $d = 9$) covariance matrix of the feature points

$$C_R(\phi) = \frac{\int_\Omega H(\phi)\left(f(x,y) - \overline{f_R(\phi)}\right)\left(f(x,y) - \overline{f_R(\phi)}\right)^T dxdy}{\int_\Omega H(\phi) dxdy}, \tag{2}$$

where $\overline{f_R(\phi)}$ is the mean of $\{f(x, y)\}_{(x,y)\in R}$,

$$\overline{f_R(\phi)} = \frac{\int_\Omega H(\phi) f(x,y) dxdy}{\int_\Omega H(\phi) dxdy}. \tag{3}$$

Here, $\phi = \phi(x, y)$ is the level set function whose zero level set represents evolving curve C, and $H(\phi)$ is the Heaviside function[18]. The background region covariance matrix C_{R^C} is obtained in the same way. The covariance matrix provides a natural way of fusing multiple features which may be correlated. Its diagonal entries of the covariance matrix represent the variance of each feature and the non-diagonal entries represent the correlations. The noise corrupting individual samples are largely filtered out with an average filter during covariance computation.

2.2 Image Energy

In this study, we want to model the appearance of an object using existing training data. Denote the training set of the appearance templates, i.e. covariance matrices, as $T_1, T_2, \cdots, T_M$. Because the covariance matrix does not lie in the Euclidean space, we introduce the Log-Euclidean operator[17] which maps the manifold space into the Euclidean space. Then, the templates can be written as

$$t_i = \beta\left(\log(T_i)\right), i = 1, 2, \cdots, M, \tag{4}$$

where $\log(\cdot)$ is the Log-Euclidean operator and $\beta(\cdot)$ is a function that stretches elements in matrix to a column vector. The whole training data can be represented as a matrix $D = [t_1, t_2, \cdots, t_M] \in \mathcal{R}^{N \times M}$ where $N = d \times d$ is the dimension of t_i and M is the number of templates. Then, for each template, we let $S_i = t_i - \bar{t}$ where $\bar{t} = \frac{1}{M}\sum_{i=1}^{M} t_i$ is the mean vector of all templates. As we know, PCA aims to perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible, and it is an unsupervised method. In our opinion, a target appearance could be reconstructed by these principal components which are the most important parts of these training data. Therefore, this set of vectors is then used to perform PCA which seeks a set of M orthonormal vectors u_k that can best describe the distribution of data. The vector u_k and scalar λ_k are the eigenvector and eigenvalue, respectively, of the covariance matrix

$$C = \frac{1}{M}\sum_{i=1}^{M} S_i S_i^T. \tag{5}$$

The associated eigenvalues allow us to rank the eigenvectors according to their usefulness to characterize the variation of the template. We choose $M^{'}$ eigenvectors with the largest eigenvalues. Since the eigenvectors seem to adequate for describing templates under controlled conditions. In practice, a smaller $M^{'} < M$ is sufficient for tracking, since accurate reconstruction of the template is not necessary. A new template t_c can be reconstructed or learned by the linear combination of the eigenvectors

$$t_c = \bar{t} + Uw, \tag{6}$$

where $U = [u_1 u_2 \cdots u_{M'}]$ consists the $M^{'}$ eigenvectors with the largest eigenvalues and w is the coefficient of each eigenvector. For a new frame, the coefficient w is unknown, so we should find the optimal solution to reconstruct the current template. In this paper we propose to minimize the distance between the candidate region covariance matrix and the current template, meanwhile maximize the distance between the background region covariance matrix and the current template. Thus, the image energy function for contour tracking can be defined as

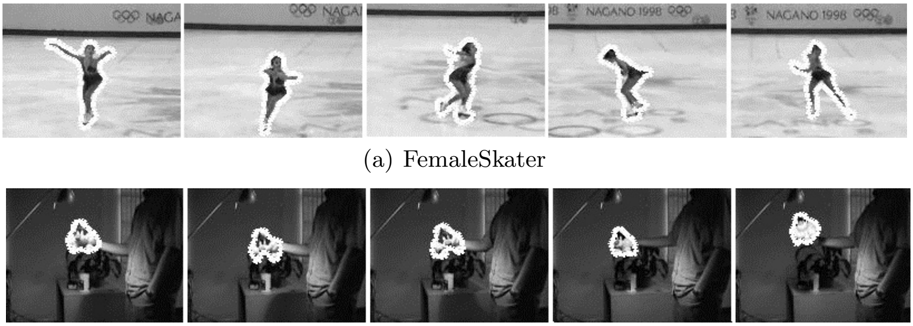
$$\begin{aligned} \rho_{im}(\phi, w) = \lambda_1 \left\|\beta\left(\log(C_R(\phi))\right) - (\bar{t} + U \cdot w)\right\|_F \\ -\lambda_2 \left\|\beta\left(\log(C_{R^C}(\phi))\right) - (\bar{t} + U \cdot w)\right\|_F, \end{aligned} \tag{7}$$

where λ_1 and λ_2 are the adjusting parameters of foreground energy and background energy respectively, and $\|\cdot\|_F$ is Frobenius norm of matrix. In order to minimize the energy function with respect to level set ϕ, we use the standard gradient descent method by solving the gradient flow equation as follows

$$\frac{\partial \phi}{\partial t} = -\frac{\partial \rho_{im}(\phi, w)}{\partial \phi}, \tag{8}$$

$$\frac{\partial w}{\partial t} = -\frac{\partial \rho_{im}(\phi, w)}{\partial w}. \tag{9}$$

For more details, please refer to [12].



(a) FemaleSkater

(b) Sylv

Fig. 1. The tracking result of our method

2.3 Shape Energy

The evolved curve may be incomplete when the image appearance cues are weak or misleading. We incorporate shape prior into the tracking algorithm to improve its robustness. Given a shape template whose level set function is $\overline{\phi}(x, y)$, the shape energy can be defined as follow

$$\rho_{sh}(\phi) = \int_{\Omega} \left(H(\phi(x, y)) - H(\bar{\phi}_{\alpha,R,T}(x, y))\right)^2 dxdy, \tag{10}$$

where $\bar{\phi}_{\alpha,R,T}(x, y)$ is a Euclidean similarity transformation of the shape template $\overline{\phi}(x, y)$. For more details of the transformation parameters, please refer to [19] and [20]. The gradient flow of shape energy can be derived as follow

$$\frac{\partial \phi}{\partial t} = -\frac{\partial \rho_{sh}(\phi)}{\partial \phi} = -2\delta(\phi)\left(H(\phi) - H(\bar{\phi}_{\alpha,R,T})\right), \tag{11}$$

where $\delta(\cdot)$ is the Dirac delta function, and $\delta(x) = \frac{d}{dx}H(x)$.

The final curve evolution equation is the combination of the image energy (10) and the shape energy (15),

$$\frac{\partial \phi}{\partial t} = -\alpha \frac{\partial \rho_{im}(\phi, w)}{\partial \phi} - \beta \frac{\partial \rho_{sh}(\phi)}{\partial \phi}, \tag{12}$$

where $\alpha > 0$ and $\beta > 0$ are fixed parameters, and α controls the image data driven force, and β controls the shape driven force. The level set ϕ will converge to object contour by solving this energy functional.

3 Experimental Results

In this section, our proposed method was tested on several image sequences which were downloaded from the Internet. The approach was implemented using

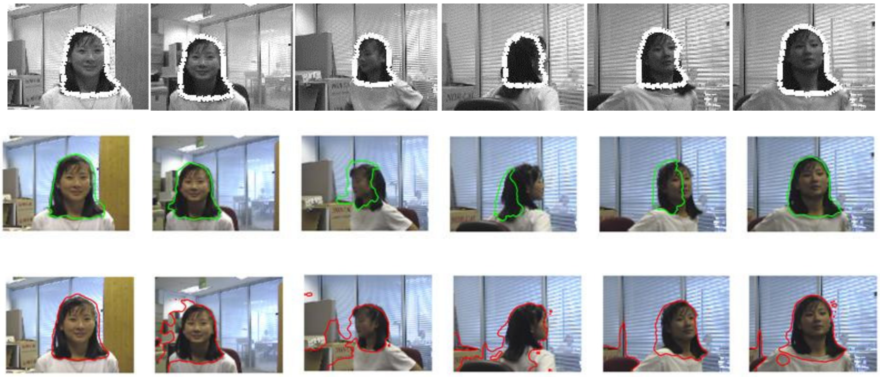


Fig. 2. The tracking results of the girl sequences. The first row shows the results of our method, and the second and the third row show the tracking results of covariance matrix matching-based tracker and distribution matching-based tracker respectively.

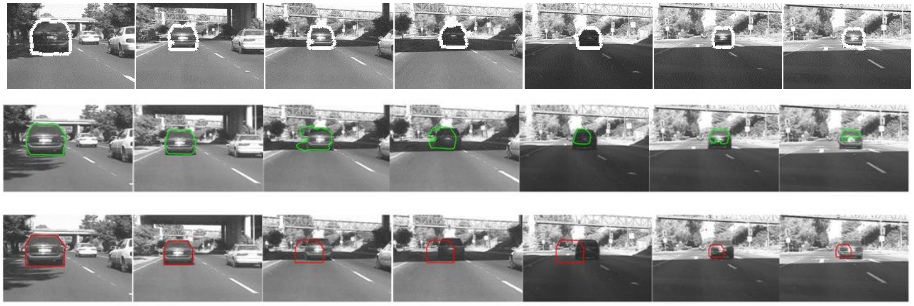


Fig. 3. The tracking results for the car sequences.The first row shows the results of our method, and the second and the third row show the tracking results of covariance matrix matching-based tracker and distribution matching-based tracker respectively.

Matlab and performed on PC with an Intel Core 2 Duo CPU (2.99GHz). During the visual tracking, the object contour in the first frame was initialized manually. The image energy item and shape energy item were normalized. In most case, the adjusting parameter λ_1 and λ_2 were set to 0.5, and the number of eigenvectors M' was set to 10.

We first test our method on the sequence FemaleSkater whose size is 320 × 240, in which the female skater changes pose over time. Fig. 1(a) shows the tracking results using our proposed method. The second image sequence, Sylv, shown in Fig. 1(b) contains a toy in different scale, lighting conditions and affine transformation. Our method can track the target curve accurately.

As a baseline, we compare our method with two other trackers. The first is covariance matching tracker with a fixed template [12] and the second is the distribution based tracker [3]. As is shown in Fig. 2, our method is able to

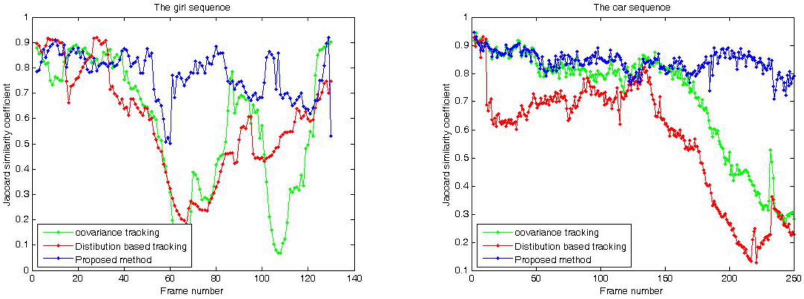


Fig. 4. Comparison of the Jaccard similarity coefficient of our method with the distribution based tracker and the covariance tracker

track the target in the girl sequence whose size is 128 × 96 undergoing gradual scale changes. Furthermore, our method is able to track the target with severe appearance changes. That is because the proposed method can efficiently learn an appearance template representation using PCA during tracking the target. In contrast, it is difficult for the covariance matching tracker with a fixed template and for the distribution based tracker.

Fig. 3 shows the tracking results of our method for the car sequence whose size is 360 × 240, in which a car is moving in different scale and lighting conditions. When the car moves into the shadow of the bridge, the appearance of the target becomes dissimilar with that in the previous frames. Compared with the results of the covariance matching tracker with a fixed template and the distribution based tracker, our method is able to track the target accurately. The appearance template learning metric contributes to the outperformance.

We use Jaccard similarity coefficient to evaluate the segmentation of each frame [21]. Jaccard similarity coefficient defines the similarity between the tracking result and the ground truth. We manually mark the ground truth and compute the coefficient for the girl sequence and car sequence, as is shown in Fig.4. We can see that the coefficient of our method is higher than the other two trackers, which proves the effectiveness of our method.

4 Conclusions

In this paper, we propose a PCA-based appearance template learning method for contour tracking. We perform this tracker on the level set framework using covariance of the visual object. Different from the previous algorithm, the appearance template is learned from a set of training data using PCA in this method. It can overcome the tracking difficulties caused by the object appearance changes, such as pose variations, illumination changes, and occlusions. Experimental results and evaluations demonstrate the high accuracy of the proposed method.

References

1. Paragios, N., Deriche, R.: Geodesic active regions and level set methods for motion estimation and tracking. CVIU 97, 259–282 (2005)
2. Cremers, D.: Dynamical statistical shape priors for level set-based tracking. PAMI 28, 1262–1273 (2006)
3. Freedman, D., Zhang, T.: Active contours for tracking distributions. IP 13, 518–526 (2004)
4. Li, C., Xu, C., Gui, C., Fox, M.D.: Level set evolution without re-initialization: a new variational formulation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 430–436. IEEE (2005)
5. Wu, Y., Ma, B., Li, P.: A variational method for contour tracking via covariance matching. Science China Information Sciences 55(11), 2635–2645 (2012)
6. Michailovich, O., Rathi, Y., Tannenbaum, A.: Image segmentation using active contours driven by the bhattacharyya gradient flow. IP 16, 2787–2801 (2007)
7. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: CVPR, vol. 2, pp. 142–149 (2000)
8. Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: CVPR, pp. 1910–1917 (2012)
9. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 589–600. Springer, Heidelberg (2006)
10. Wu, Y., Cheng, J., Wang, J., Lu, H.: Real-time visual tracking via incremental covariance tensor learning. In: ICCV, pp. 1631–1638 (2009)
11. Ma, B., Wu, Y., Li, P.: Level set segmentation using image second order statistics. In: SPIE, vol. 8003 (2011)
12. Ma, B., Wu, Y.: Covariance matching for pde-based contour tracking. In: ICIG, pp. 720–725 (2011)
13. Zhang, S., Zhan, Y., Dewan, M., Huang, J., Metaxas, D., Zhou, X.: Sparse shape composition: A new framework for shape prior modeling. In: CVPR, pp. 1025–1032 (2011)
14. Ross, D., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. IJCV 77, 125–141 (2008)
15. Li, Y.: On incremental and robust subspace learning. PR 37, 1509–1518 (2004)
16. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
17. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine 56, 411–421 (2006)
18. Abramowitz, M., Stegun, I.A.: Handbook of mathematical functions with formulas, graphs, and mathematical tables. National Bureau of Standards Applied Mathematics Series, vol. 55. U.S. Government Printing Office, Washington, D.C. (1964)
19. Dryden, I., Mardia, K.: Statistical shape analysis, vol. 4. John Wiley & Sons New York (1998)
20. Zhang, T., Freedman, D.: Tracking objects using density matching and shape priors. In: ICCV, pp. 1056–1062 (2003)
21. Udupa, J., Leblanc, V., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L., Hirsch, B., Woodburn, J.: A framework for evaluating image segmentation algorithms. CMIG 30, 75–87 (2006)

A Novel Variational PDE Technique for Image Denoising

Tudor Barbu

Institute of Computer Science of the Romanian Academy, Iaşi, Romania
tudbar@iit.tuiasi.ro

Abstract. A robust variational PDE model for image noise removal is proposed in this paper. One considers an energy functional to be minimized, based on a novel smoothing constraint. Then, the corresponding Euler-Lagrange equation is determined. The obtained PDE model is solved, by using a numerical discretization scheme. Some results of our image denoising experiments and method comparisons are also described in this article.

Keywords: image denoising, variational technique, PDE models, Euler Lagrange equations, energy functional minimization, smoothing function, discretization scheme.

1 Introduction

During the past two decades, the mathematical models have been increasingly used in some traditionally engineering domains like signal and image processing, analysis, and computer vision [1]. The variational and Partial Differential Equation (PDE) based techniques have been widely used and studied in this fields in the past few years because of their modelling flexibility and some advantages of their numerical implementation [2].

Thus, some important application areas of the variational PDE methods are image denoising, image reconstruction (inpainting), image segmentation (contour tracking), image registration and optical flow [1, 2]. We consider a variational approach for image denoising in this paper.

Image noise removal with feature preservation is still a focus in the image processing area and serious challenge for the researchers. An efficient image denoising approach must not only substantially reduce the noise amount but also preserve the boundaries and other characteristics [3]. Conventional image filters, like averaging, median, or the classic 2D Gaussian filter succeed in noise reduction, but also have an edge-blurring effect [4].

The linear PDE-based denoising techniques are derived from the the use of the Gaussian filter in multiscale image analysis [4, 5]. The convolution of an image with a 2D Gaussian kernel amounts to solve the diffusion equation in two dimensions (heat equation). The nonlinear PDE-based approaches are able to smooth the images while preserving their edges, also avoiding the localization problems of linear filtering. The most popular nonlinear PDE denoising method is the influential nonlinear anisotropic

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 501–508, 2013.

diffusion scheme developed by P. Perona and J. Malik in 1987 [5, 6]. Numerous denoising techniques derived from their algorithm have been proposed since then [5].

There are many ways to get the nonlinear PDEs. In image processing and computer vision it is very common to obtain them from some variational problems. The basic idea of any variational PDE technique is the minimization of an energy functional [1, 7-9]. The variational techniques have important advantages in both theory and computation, compared with other methods. They can achieve high speed, accuracy, and stability using the extensive results of the numerical PDE approaches.

An influential variational denoising and restoration model was developed by Rudin, Osher and Fetami in 1992. Their technique, named Total Variation (TV) denoising, is based on the minimization of the TV norm [7]. TV denoising is remarkably effective at simultaneously preserving boundaries whilst smoothing away noise in flat regions, but it also suffers from the *staircasing* effect and its corresponding Euler-Lagrange equation is highly nonlinear and difficult to compute. In recent years, many PDE approaches that improve this classical variational model have been proposed [1].

The novel PDE variational technique provided in this paper achieves an efficient smoothing result while preserving the image edges and also solves the staircase problem [8, 9, 12]. The main contribution of our denoising variational model is the robust smoothness term (regularizer) introduced in the energy functional that is described in the next section. Also, we provide a satisfactory discretization of the PDE model, a good approximation of the Euler-Lagrange equation being described in the third section of this article.

Numerous image denoising experiments using this method and method comparisons have been performed. They are discussed in the fourth section. The conclusions are presented in the last section and the paper ends with a list of references.

2 Variational Model for Image Noise Reduction

The general variational framework used in image processing and computer vision is characterized by an energy functional having the following form:

$$E\left[u(x)\right] = \int_{\Omega} \left(D(u) + S(u)\right) dx \tag{1}$$

where $D(u)$ represents the *data component* and $S(u)$ is the *smoothing term* of the functional [7, 10]. So, one must determine the unknown function $u(x)$ on the domain $\Omega \subset R^2$, that minimizes the above energy:

$$u_{\min} = \arg \min_{u \in U} E\left[u(x)\right] \tag{2}$$

In the variational image denoising case, one considers an image u_0 affected by Gaussian noise. The general form of the energy functional used by variational image smoothing processes is:

$$E[u] = \int_{\Omega} (u - u_0)^2 + \alpha\psi\left(\|\nabla u\|^2\right), \quad \alpha > 0 \tag{3}$$

where the function ψ represents the regularizer (penalizer) of the smoothing term and α is the regularization parameter or smoothness weight [10].

We develop an efficient smoothing component, based on a novel penalizer function and a proper value of the smoothness weight. Thus, we consider the following regularizer: $\psi : [0, \infty) \rightarrow [0, \infty)$:

$$\psi(s) = \eta\sqrt{\frac{k}{\beta}} \ln\left(s + \sqrt{s^2 + \frac{\gamma}{\beta}}\right) + \nu \cdot s; \quad k > 0, \eta, \beta, \gamma, \nu \in (0,1) \tag{4}$$

We consider some proper values for the penalizer's parameters. The values of $k, \eta, \beta, \gamma, \nu$ and α which provide a successful denoising are specified in section 4, related to numerical experiments. Then, we compute a minimizer for the energy functional given by (3), using the function ψ given by (4):

$$u_{\min} = \arg\min_{u \in U} E(u) = \arg\min_{u \in U} \int_{\Omega} (u - u_0)^2 + \alpha\psi\left(\|\nabla u\|^2\right) dxdy \tag{5}$$

The minimization result $u_{\min}$ will correspond to the denoised (smoothed) image. The minimization process is performed by solving the following Euler-Lagrange equation [7,10,11]:

$$u - u_0 - \alpha div\left(\psi'\left(\|\nabla u\|^2\right)\nabla u\right) = 0 \Leftrightarrow \frac{u - u_0}{\alpha} - div\left(\psi'\left(\|\nabla u\|^2\right)\nabla u\right) = 0 \tag{6}$$

Thus, we obtain the following PDE equation:

$$\frac{\partial u}{\partial t} = div\left(\psi'\left(\|\nabla u\|^2\right)\nabla u\right) - \frac{u - u_0}{\alpha} \tag{7}$$

where the positive function ψ' is obtained by computing the derivative of the function given by (4) as follows:

$$\psi'\left(s^2\right) = \frac{\nu\sqrt{\beta s^2 + \gamma} + \eta\sqrt{k}}{\sqrt{\beta s^2 + \gamma}} \tag{8}$$

Therefore, the partial differential equation (7) becomes

$$\begin{cases} \dfrac{\partial u}{\partial t} = div\left(\dfrac{\eta\sqrt{\beta\|\nabla u\|^2 + \gamma} + \alpha\sqrt{k}}{\sqrt{\beta\|\nabla u\|^2 + \gamma}} \cdot \nabla u \right) - \dfrac{u - u_0}{\alpha} \\ \\ u(0, x, y) = u_0 \end{cases} \qquad (9)$$

One can demonstrate the PDE model given by (9) converges to a unique strong solution, that is $u^* = u_{\min}$. We propose a robust discretization scheme for solving it, which is described in the next section.

3 Discretization Scheme for the PDE Model

We consider a proper numerical approximation of the proposed PDE model's solution. Thus, our discretization scheme uses a 4-NN discretization of the Laplacian operator [6].

From (7) we have $\frac{\partial u}{\partial t} = div\left(\psi'\left(\|\nabla u\|^2\right)\nabla u\right) - \frac{u - u_0}{\alpha}$, which leads to the following relation:

$$u(x, y, t+1) \cong u(x, y, t) + div\left(\psi'\left(\|\nabla u\|^2\right)\nabla u\right) - \frac{u - u_0}{\alpha} \qquad (10)$$

One can approximate (10) using the image gradient magnitudes in particular directions, as following:

$$u^{t+1} = u^t + \lambda \sum_{q \in N(p)} \psi'\left(\left\|\nabla u_{p,q}(t)\right\|^2\right)\nabla u_{p,q}(t) - \frac{u - u_0}{\alpha} \qquad (11)$$

where $\lambda \in (0,1)$ and $t = 1, \ldots, N$.

In the equation above $N(p)$ represents the the 4-neighborhood of the argument pixel, described by its coordinates, $p = (x, y)$. Obviously, it represents a set of image pixels given by their coordinates:

$$N(p) = \{(x-1, y), (x+1, y), (x, y-1), (x, y+1)\} \qquad (12)$$

Also, $\nabla u_{p,q}$ is the image gradient magnitude in the direction given by pixel q at iteration t, being computed as follows:

$$\nabla u_{p,q}(t) = u(q, t) - u(p, t) \qquad (13)$$

The maximum number of iterations, *N*, is empirically chosen. The proposed iterative denoising scheme applies the operation given by (11) for each *t* value, from 0 to *N*. Our noise removal technique produces the smoothed image u^N from the noised image $u^0 = u_0$ in a relatively small number of steps, being characterized by a quite low *N* value.

That means, the PDE model developed here converges fast to the solution $u^N \cong u_{\min}$. The effectiveness of the proposed PDE denoising approach and its discretization is proved by the satisfactory image smoothing results obtained from our experiments. These numerical experiments are discussed in the next section of the article.

4 Experiments and Method Comparisons

The described variational PDE denoising technique have been applied on numerous image datasets. We have performed numerous image smoothing experiments, using the proposed technique, on hundreds noisy images, and obtained very good results.

Thus, the original images have been corrupted with various level of Gaussian noise (various values for mean and variance). Then, the denoising model have been applied to them with some properly chosen parameters which provide best results. These empirically detected parameter values are:

$$\alpha = 9, k = 25, \eta = 0.7, \beta = 0.66, \gamma = 0.5, \nu = 0.2, \lambda = 0.3, N = 15 \quad (14)$$

We assess the performance of our noise reduction method using the *norm of the error image* measure [8, 9]. Thus, if u_{orig} represents the original (noise-free) form of the image, then the norm of the error image is computed as:

$$NE(u) = \sqrt{\sum_{x=1}^{X} \sum_{y=1}^{Y} (u^N(x, y) - u_{orig}(x, y))^2} \quad (15)$$

where $[X \times Y]$ is the image dimension. Our denoising techniques provides low enough values for this performance measure.

From the performed method comparisons we have found that our variational technique outperforms other noise removal approaches. Thus, we have compared it with some other PDE-based methods and also with some non-PDE denoising algorithms. Our approach provides considerable better image denoising and edge-preserving results than non-PDE image filters, like Gaussian, average and median filters. It also achieves a better smoothing and, given its lower time complexity, converges faster than other variational schemes, such as the quadratic variational model, characterized by a regularizer $\psi(s^2) = s^2$, or the Perona-Malik variational scheme, given by

$\psi(s^2) = \lambda^2\left(\log\left(1 + \frac{s^2}{\lambda^2}\right)\right)$ [10]. Because of its low execution time, this method can be used for denoising large image sets, like those of social networks [12].

Several image denoising results and method comparisons are described in the next figures and tables. In Fig.1, there are displayed: a) the original $[512\times512]$ *Lena* image; b) the image corrupted with Gaussian noise given by $\mu = 0.211$ and *var* = 0.023; c) the image smoothed by our variational model; d) quadratic denoising; e) Perona-Malik noise removal; f) – i) denoising results achieved by the 2D Gaussian, average, median and Wiener $[3\times3]$ filter kernels. The corresponding norm of the error values are displayed in Table 1.

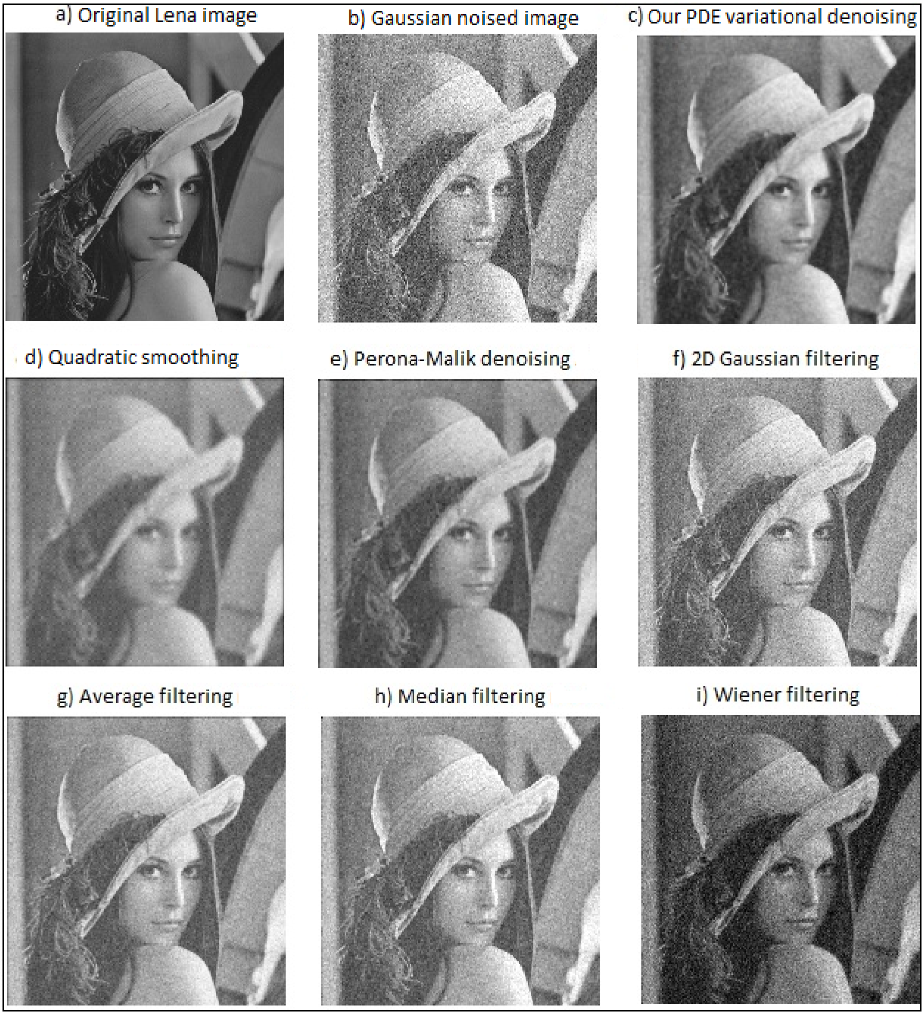


Fig. 1. Lena image denoised using various smoothing techniques

Table 1. Norm-of-the-error values for several noise removal techniques

Our alg.	Quadratic	P-M	Gaussian	Average	Median	Wiener
4.9×10^{3}	6.2×10^{3}	5.9×10^{3}	7.4×10^{3}	6.4×10^{3}	6×10^{3}	5.6×10^{3}

In Fig. 2 the same denoising models are applied on the *Baboon* image, while the corresponding values of the *NE* measure are registered in Table 2. As one can see in Fig. 1 and Fig. 2, the variational approach proposed here provides the best edge-preserving image smoothing. The staircasing effect, representing creation in the image of flat regions separated by artifact boundaries [9, 13], is also removed by our denoising technique.

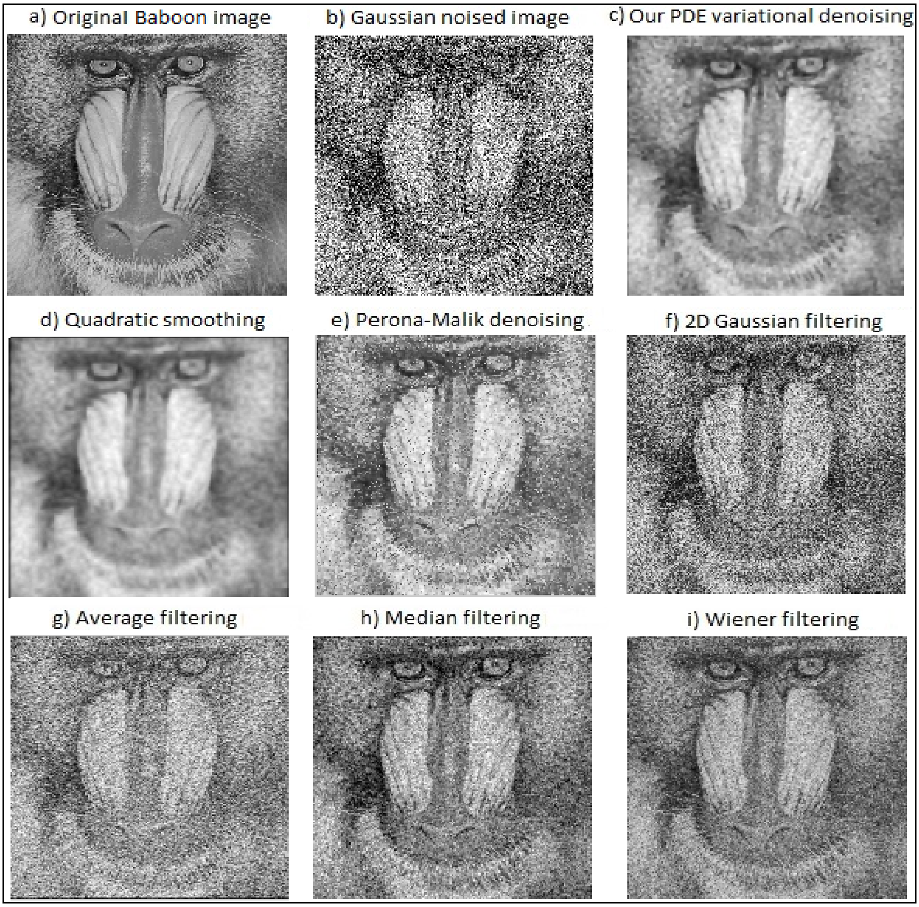

a) Original Baboon image
b) Gaussian noised image
c) Our PDE variational denoising
d) Quadratic smoothing
e) Perona-Malik denoising
f) 2D Gaussian filtering
g) Average filtering
h) Median filtering
i) Wiener filtering

Fig. 2. Baboon image denoised using various smoothing techniques

Table 2. Norm-of-the-error values for several noise removal techniques

Our alg.	Quadratic	P-M	Gaussian	Average	Median	Wiener
5×10^{3}	6×10^{3}	5.9×10^{3}	7.3×10^{3}	6.5×10^{3}	6.1×10^{3}	5.8×10^{3}

5 Conclusions

We have proposed a variational PDE denoising approach in this paper. This technique performs an efficient noise removal and also preserves the boundaries of the image.

The main original contribution of this article is the efficient smoothing component introduced in the energy functional of the variational model. It is based on a novel regularizer function. Also, we propose a robust discretization of the PDE model given by the corresponding Euler-Lagrange equation. Our developed variational technique reduces also the staircasing effects and converges fast to the solution represented by the denoised image. It also outperforms many other variational PDE methods and non-PDE denoising techniques [1,13], as resulting from the performed experiments and the method comparison.

We intend to further investigate this variational scheme and to provide more mathematical treatment of it in the future. Thus, the demonstration of the convergence of our PDE model to a unique strong solution will be the subject of our future work in this domain. PDE-based color image denoising [14] will also be a next research field.

References

1. Chan, T., Shen, J., Vese, L.: Variational PDE Models in Image Processing. Notices of the AMS 50(1) (2003)
2. Song, B.: Topics in Variational PDE Image Segmentation, Inpainting and Denoising, University of California (2003)
3. Ning, H.E., Ke, L.U.A.: A Non Local Feature-Preserving Strategy for Image Denoising. Chinese Journal of Electronics 21(4) (2012)
4. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall, NJ (1989)
5. Weickert, J.: Anisotropic Diffusion in Image Processing. European Consortium for Mathematics in Industry. B. G. Teubner, Stuttgart (1998)
6. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. In: Proc. of IEEE Computer Society Workshop on Computer Vision, pp. 16–22 (November 1987)
7. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1), 259–268 (1992)
8. Barbu, T., Barbu, V., Biga, V., Coca, D.: A PDE variational approach to image denoising and restorations. Nonlinear Anal. RWA 10, 1351–1361 (2009)
9. Barbu, T.: Variational Image Denoising Approach with Diffusion Porous Media Flow. Abstract and Applied Analysis 2013, Article ID 856876, 8 pages (2013)
10. Popuri, K.: Introduction to Variational Methods in Imaging. In: CRV 2010 Tutorial Day (2010)
11. Fox, C.: An introduction to the calculus of variations. Courier Dover Publications (1987)
12. Alboaie, L., Vaida, M.F.: Trust and Reputation Model for Various Online Communities. Studies in Informatics and Control 20(2), 143–156 (2011)
13. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Model. Simul. 4(2), 490–530 (2005)
14. Kim, S.: PDE-based image restoration: A hybrid model and color image denoising. IEEE Transactions on Image Processing 15(5), 1163–1170 (2006)

Coronary Plaque Boundary Calculation in IVUS Image by Modified PMD Filter and Fuzzy Inference

Syaiful Anam[1,2], Eiji Uchino[1,3], and Noriaki Suetake[1]

[1] Graduate School of Science and Engineering, Yamaguchi University,
1677-1 Yoshida, Yamaguchi 753-8512, Japan
{r501wa,uchino,nsuetake}@yamaguchi-u.ac.jp
http://www.ic.sci.yamaguchi-u.ac.jp

[2] Mathematics Department, University of Brawijaya,
Veteran St., Malang 65145, Indonesia
syaiful@ub.ac.id
http://www.ub.ac.id

[3] Fuzzy Logic Systems Institute,
680-41 Kawazu, Iizuka, Fukuoka 820-0067, Japan
http://flsi.cird.or.jp

Abstract. In this paper, we propose a method for coronary plaque boundary calculation in Intravascular Ultrasound (IVUS) image by using a modified Perona Malik Diffusion (PMD) filter and the Takagi-Sugeno (T-S) fuzzy model. The modified PMD filter is designed based on the coronary plaque boundary direction in IVUS image to reduce the speckle noise and to enhance the coronary plaque boundary. Searching areas for the plaque boundaries are automatically set by using the weighted image separability and some heuristic rules. The coronary plaque boundaries are interpolated by the polynomials inferred by the fuzzy rules. It has been confirmed that the accuracy of the proposed method is better than that of the method using the normal PMD filter.

Keywords: Coronary plaque boundary calculation, IVUS image, modified PMD filter, T-S fuzzy model.

1 Introduction

Acute Coronary Syndromes (ACS) is one of the leading hospitalizations which is caused by a rupture of vulnerable plaque. The plaques are built up inside the coronary arteries which cause heart attack.

One of the medical imaging methods to diagnose ACS is Intravascular Ultrasound (IVUS) method [1]. It provides a real time cross-sectional image of a coronary artery *in vivo*. IVUS image is used for the inner and outer coronary plaque boundaries calculation to evaluate the quantitative assessment of the coronary plaque compositions.

The coronary plaque boundaries calculation task is very hard for medical doctors, because currently those boundaries are manually drawn by them. In

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 509–516, 2013.

order to reduce the workload of the medical doctors, an automatic coronary plaque boundaries calculation method with high accuracy is strongly desired.

For this goal, in our previous works [2] and [3], the fuzzy inference based method was proposed and applied for this problem. Fuzzy inference model employed is Takagi-Sugeno (T-S) fuzzy model [4]. In [2], Membership Functions (MSFs) in the antecedent parts of the fuzzy rules were allocated adaptively. In [3], weighted image separability and heuristic rules were used for determining the seed points automatically.

The plaque boundary in IVUS image is very difficult to recognize because of heavy speckle noise. For this reason, preprocessing of IVUS image is strongly required. In [2] and [3], Perona-Malik Diffusion (PMD) filter [5] is used to reduce the speckle noise. The PMD filter is a method to preserve an image edge effectively.

When the PMD filter is applied to an IVUS image, it not only reduces the speckle noise but also enhances the image edges. However, the diffusion direction and its strength are very important factors to enhance the image edges and to reduce the speckle noise. In the previous methods [2] and [3], the numbers of direction was four, and the diffusion strength was set to the same value in all directions. Neither, the plaque boundary direction in IVUS image was considered.

In this paper, we propose a modified diffusion direction and strength in the PMD filter. The effectiveness of the proposed method is verified through the experiments using the real IVUS images.

2 Coronary Plaque Boundary Calculation in IVUS Image

The IVUS method is one of the applications of ultrasound technology which has many applications in medical diagnosis. The IVUS method is a very ingenious method to observe from inside the blood vessel out through the surrounding blood column, visualizing the coronary plaque *in vivo*.

The image shown in Fig. 1 is called a “B-mode image,” which is constructed of the amplitude information of the received ultrasound Radio Frequency (RF) signals. The sampled RF signals are transformed into intensities, and the intensities in all radial directions are used to form a tomographic cross-sectional image of a coronary artery.

The boundaries of plaque need to be calculated for the diagnosis of plaque. Fig. 1(a) shows two boundaries of plaque, i.e., Luminal Boundary (LB) and Adventitial Boundary (AB). In [2] and [3], the plaque boundaries are approximated by the piecewise polynomials inferred by the Takagi-Sugeno (T-S) fuzzy model based on the given seed points.

2.1 Image Separability

The image separability [6] is used to obtain a candidate of plaque boundary. The area with high separability becomes a candidate of plaque boundary.

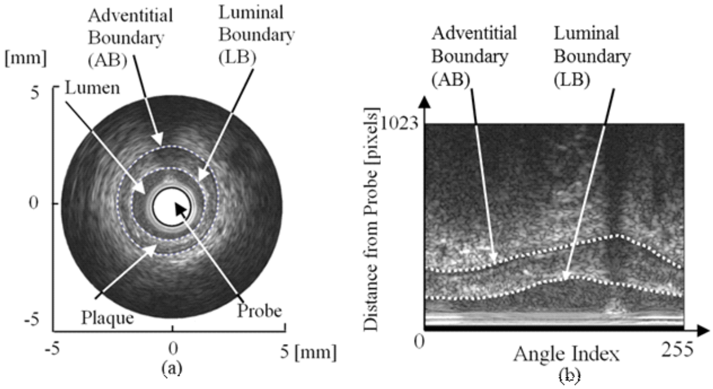


Fig. 1. IVUS B-mode image. (a) B-mode image in the Cartesian coordinates. (b) B-mode image in the polar coordinates.

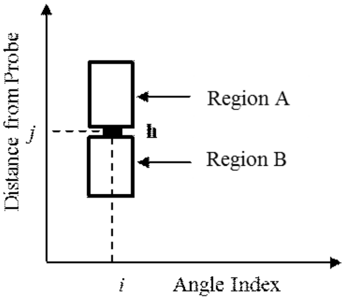


Fig. 2. Calculation of the image separability

The weighted image separability for pixel $\mathbf{h} = (i, j)$ shown in Fig. 2 is defined by:

$$\eta_{\mathbf{h}}^{w} = \eta_{\mathbf{h}}(\frac{I_{max} - \overline{I}_A}{I_{max}} \times \frac{\overline{I}_B}{I_{max}})^2, \tag{1}$$

where I_{max} is a maximum intensity on the whole IVUS image. $\eta_{\mathbf{h}}$ is an image separability for pixel $\mathbf{h}$. $\overline{I}_A$ and $\overline{I}_B$ represent the averages of intensities in the regions of A and B, respectively.

2.2 Takagi-Sugeno (T-S) Fuzzy Model for Plaque Boundary Calculation

The plaque boundary is inferred by the T-S fuzzy model. The boundary is piecewise approximated by the series of the following fuzzy rules:

$$\text{IF} \quad \mathbf{x}_i \quad \text{is} \quad \mathbf{A}_u \quad \text{THEN} \quad f_u(\mathbf{x}_i) = \mathbf{a}_u \mathbf{x}_i + \mathbf{b}_u, \tag{2}$$

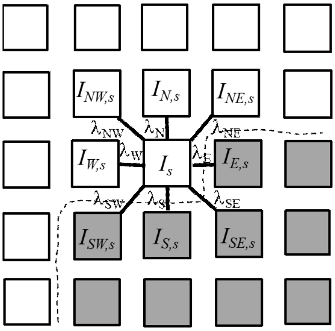


Fig. 3. Diffusion directions of the modified PMD filter

where $\mathbf{A}_u$ is a fuzzy set with the Membership Function (MSF) $\mu_u(\mathbf{x}_i)$. $\mathbf{x}_i$ corresponds to the angle index, and $f_u(\mathbf{x}_i)$ is a linear function.

In the antecedent part of the fuzzy rule, the complementary triangular MSFs are used. The u-th rule thus stands for a piecewise approximation of the plaque boundary by a linear function in the interval $[z_u, z_{u+1}]$. The inferred boundary is given by:

$$\hat{y}_i(\mathbf{x}_i) = \mu_u(\mathbf{x}_i) f_u(\mathbf{x}_i) + \mu_{u+1}(\mathbf{x}_i) f_{u+1}(\mathbf{x}_i). \tag{3}$$

The optimum coefficients in the consequent part of the fuzzy rule are determined with use of Weighted Least Square Method (WSLM) so as to minimize the following weighted error criterion:

$$E = \sum_{j=0}^{J-1} \sum_{i=0}^{I-1} \eta_{\mathbf{h}}^{w} (y_i - \hat{y}_i(\mathbf{x}_i)), \tag{4}$$

where $\eta_{\mathbf{h}}^{w}$ is a weighted image separability of pixel $\mathbf{h} = (i, j)$. In this method, $\eta_{\mathbf{h}}^{w}$ inside the search area are used as the weights of WLSM [2] [3].

3 Proposed Method

We propose a modified direction diffusion in the PMD filter based on the direction of the plaque boundary of the IVUS image.

3.1 Modified PMD Filter

The anisotropic diffusion filter was originally proposed by Perona and Malik [5] in order to realize an edge-preserved smoothing of image. The discrete version of PMD diffusion process is defined as follows:

$$I_s^{(n+1)} = I_{\mathbf{s}}^{(n)} + \frac{\lambda}{|\phi_s|} \sum g(\nabla I_{s,p}^{(n)}) I_{\mathbf{s},p}^{(n)}, \tag{5}$$

where $\mathbf{s}(x, y)$ and p are the coordinates of the pixel of concern and its neighboring pixels, respectively. $I_s^{(n)}$ is an intensity at $\mathbf{s}$ with an iteration count n. $|\phi_s|$ represents the number of diffusion directions. λ is a parameter.

$g(\cdot)$ refers to an edge stopping function, which is a decreasing function of the gradient of image. $g(\cdot)$ takes large values at the regions where the intensity gradients are low. On the contrary, it takes small values at the regions where the intensity gradients are high.

The PMD filter enhances the plaque boundary when the direction and strength of diffusion are correct. So that the direction and strength of diffusion are very important factors to enhance the edge of image and to reduce the noise. If the strength of diffusion is too large, the edge of image tends to be lost. On the contrary, if the strength of diffusion is too small, the noise of image cannot be reduced.

We propose here the modified direction and strength of diffusion of the PMD filter considering the plaque boundary direction in IVUS image. In Fig. 1(b), we can observe the boundaries of plaque in horizontal direction. It means that in order to preserve the plaque boundaries, the diffusion strength in horizontal direction should be smaller than that in other directions. Because of this we propose a new structure for diffusion directions. Fig. 3 shows the diffusion directions of the modified PMD filter.

The modified PMD filter has 8 directions and 8 different parameters in each direction. It means that the particles in the modified PMD filter move in 8 directions with different strength in each direction. By modifying the original PMD of (5) based on the diffusion direction in Fig. 3, the proposed iteration formula for diffusion process of the modified PMD filter is given as follows:

$$I_s^{(n+1)} = I_{\mathbf{s}}^{(n)} + \frac{1}{|\phi_s|} \sum_k \lambda_k g(\nabla I_{k,\mathbf{s}}^{(n)}) I_{k,\mathbf{s}}^{(n)}, \tag{6}$$

where $k = \{NW, N, NE, E, SE, S, SW, W\}$. They are the eight neighboring pixels in North West, North, North East, East, South East, South, South West and West.

3.2 Experimental Results and Discussions

In the experiments, we used three IVUS images, and the proposed method was compared with the method with the normal PMD filter in [3]. In the proposed method and the method with the normal PMD filter [3], the seed points were automatically placed as in [3].

After doing several experiments, the parameters of the diffusion filter of (6) were set as $\lambda_N = \lambda_S = 1$ and $\lambda_W = \lambda_E = \lambda_{NW} = \lambda_{SW} = \lambda_{SE} = \lambda_{NE} = 1.2$. In the experiments, the maximum iteration number of diffusion process was set to 3,000. If the diffusion parameters are too large, the plaque boundary will be lost. But, if the diffusion parameters are too small, the noise can not be reduced.

Fig. 4(a) shows the result by the method with the normal PMD filter [3]. The result by the modified PMD filter is shown in Fig. 4(b). It can be observed that

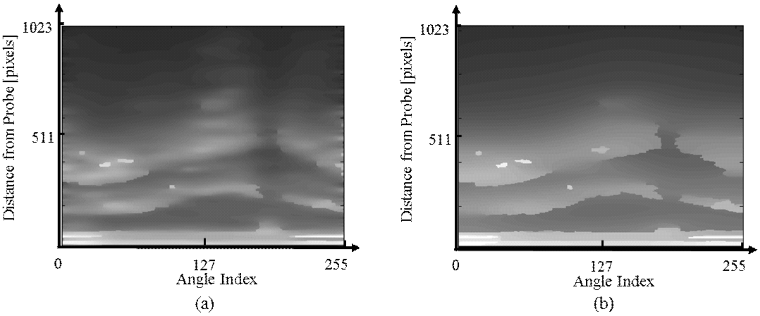


Fig. 4. Diffusion filter results for image 1. (a) The method with the normal PMD filter [3]. (b) The proposed method.

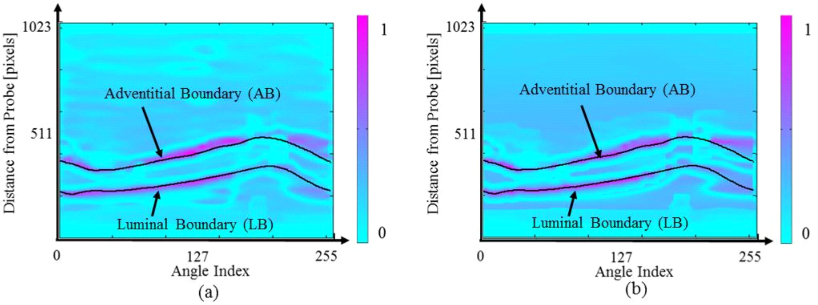

Distance from Probe [pixels]
1023
511
Adventitial Boundary (AB)
Luminal Boundary (LB)
1
0
0
127
255
Angle Index
(a)
Distance from Probe [pixels]
1023
511
Adventitial Boundary (AB)
Luminal Boundary (LB)
1
0
0
127
255
Angle Index
(b)

Fig. 5. Weighted image separability for image 3. (a) The method with the normal PMD filter [3]. (b) The proposed method.

the plaque boundaries by the modified PMD filter are clearer than those by the method with the normal PMD filter [3].

Fig. 5 shows the weighted image separability after applying the method in [3] (Fig. 5(a)) and the proposed method (Fig. 5(b)). It is seen that the desired (true) boundaries are located nearer to the center of the area with high weighted image separability by the proposed method than by the method in [3].

Table 1 shows the Root Mean Square Errors (RMSEs) between the desired and the calculated plaque boundaries. The desired boundaries were decided empirically by the experts based on the difference of image brightness. The RMSEs of the proposed method are smaller than those of the method in [3] for the most parts of LB and all parts of AB.

Fig. 6 shows the comparisons of the plaque boundary calculation by the method in [3] and the proposed method for image 1. It is seen that the calculated boundary by the proposed method (green line) is closer to the desired boundary (red line) than that by the method in [3] (blue line). The proposed method has better performance.

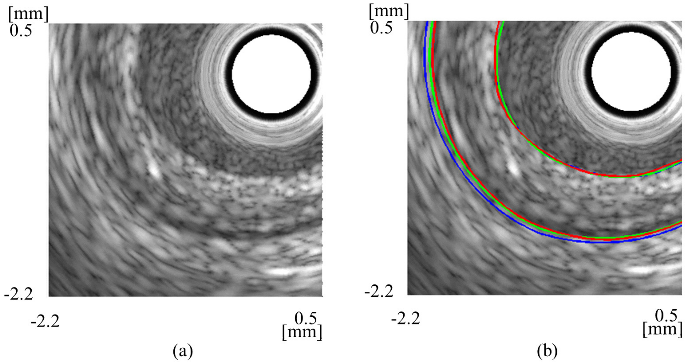


Fig. 6. Comparison of the plaque boundary calculation. (a) IVUS image to be processed. (b) Boundary calculation results. Blue and green lines indicate the calculated boundaries by the method with the normal PMD filter [3] and the proposed method, respectively. Red line is the desired boundary.

Table 1. RMSEs of boundary extraction results

Method	Num. of	Image 1		Image 2		Image 3	
	MSFs	LB	AB	LB	AB	LB	AB
Method with Normal PMD Filter[3]	Auto	13.7	28.4	**20.0**	30.2	28.1	35.4
Proposed Method	Auto	10.6	34.9	20.3	14.0	14.0	36.1
	3	9.5	15.0	20.3	14.0	14.0	34.6
	4	12.2	28.5	19.7	22.1	9.7	34.6
	Average	**10.8**	**26.1**	20.1	**16.7**	**12.6**	**35.1**

4 Conclusion

We have proposed a method for coronary plaque calculation in the IVUS image by using the modified PMD filter and the Takagi-Sugeno fuzzy model. The proposed method has better extraction performance than ever in terms of the calculation accuracy.

Acknowledgments. This work was supported by the Grant-in-Aid for Scientific Research (B) of the Japan Society for Promotion of Science (JSPS) under the Contract No. 23300086. Syaiful Anam would also like to thank DGHE postgraduate scholarship of Indonesia for supporting his stay at Yamaguchi University, Japan.

References

1. Potkin, B.N., Bartorelli, A.L., Gessert, J.M., Neville, R.F., Almagor, Y., Roberts, W.C., Leon, M.B.: Coronary Artery Imaging with Intravascular High-frequency Ultrasound. J. Circulation 81, 1575–1585 (1990)

2. Uchino, E., Suetake, N., Koga, T., Ichiyama, S., Hashimoto, G., Hiro, T., Matsuzaki, M.: Automatic Plaque Boundary Extraction in Intravascular Ultrasound Image by Fuzzy Inference with Adaptively Allocated Membership Functions. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008, Part II. LNCS, vol. 5507, pp. 583–590. Springer, Heidelberg (2009)
3. Koga, T., Ichiyama, S., Uchino, E., Suetake, N., Hiro, T., Matsuzaki, M.: Fully Automatic Boundary Extraction of Coronary Plaque in IVUS Image by Anisotropic Diffusion and T-S Type Fuzzy Inference. In: Gao, X.-Z., Gaspar-Cunha, A., Köppen, M., Schaefer, G., Wang, J. (eds.) Soft Computing in Industrial Applications. AISC, vol. 75, pp. 139–147. Springer, Heidelberg (2010)
4. Takagi, T., Sugeno, M.: Fuzzy Identification of Systems and Its Applications to Modeling and Control. IEEE Transactions on Systems, Man, and Cybernetics SMC–15, 116–132 (1985)
5. Perona, P., Malik, J.: Scale-Space and Edge Detection Using Anistropic Diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 629–639 (1990)
6. Fukui, K.: Edge Extraction Method Based on Separability of Image Features. IEEE Transactions on Information Systems E78-D, 1533–1538 (1995)

Multiple Kernel Learning with Hierarchical Feature Representations

Juhyeon Lee, Jae Hyun Lim, Hyungwon Choi, and Dae-Shik Kim

Department of Electrical Engineering, Korea Advanced Institute
of Science and Technology, 291 Daehak-ro, Daejeon 305-701, South Korea
{jhlee89,lim-0606,hyungwon.choi}@kaist.ac.kr,
dskim@ee.kaist.ac.kr

Abstract. In this paper, we suggest multiple kernel learning with hierarchical feature representations. Recently, deep learning represents excellent performance to extract hierarchical feature representations in unsupervised manner. However, since fine-tuning step of deep learning only considers global level of features for classification problems, it makes each layers hierarchical features intractable. Therefore, we propose a method to employ the combined multiple levels of pre-trained features via Multiple Kernel Learning (MKL). MKL is lately proposed optimization problem in classification and is applied to various machine learning problems. MKL automatically finds the best combination of kernels. By applying multiple kernel learning to hierarchical features pre-trained by deep learning, we obtain the optimal combinations of multiple levels of features for the classification task. Also, MKL is applied to analyze the contribution of each layer of features for classification by obtained weight of each kernel.

Keywords: Multiple Kernel Learning, Deep Learning, Deep Belief Network.

1 Introduction

Recently, deep learning is reported as an effective way to extract feature representations in unsupervised manner [1], [2]. More specifically, deep learning algorithms learn feature hierarchies from lower level to those of higher level; higher level features are detected over the previous level. In general, there are two steps in deep learning. First, features of each layer are pre-trained with layer-wise unsupervised learning, and these pre-trained features are then fine-tuned throughout whole layers by supervised learning. [3] Like multi-layer perceptron, the whole-layer fine-tuning only utilizes global level features. Therefore, features from different hierarchies are intractable for classification problems. Accordingly, instead of employing fine-tuning step, we use integrated multiple levels of pre-trained features in order to decide for classification problems collectively. Moreover, the previous study [4] suggests that globally trained features from each layer are concatenated into single vector representations and this feature vector

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 517–524, 2013.

is classified by Support Vector Machine. The combined features in different hierarchies show better performance in object recognition problem. However, those simple concatenations have limitations as well in mathematical sense; the concatenation of different types of information in un-normalized way diminishes the information.

Throughout this paper, we propose a way to employ combined hierarchical feature representations via Multiple Kernel Learning, which provides a new approach to use pre-trained features.

Multiple Kernel Learning (MKL) [5] is recently proposed optimization problem in classification and is applied to various machine learning tasks such as object classification [6], object detection [7] and bioinformatics [8]. Unlike SVM, MKL learns the optimized combination of multiple kernels to concatenate features spaces and data from different sources and maximize margin as well. In other words, for each machine learning problem MKL automatically finds out the best kernel combination. Previous MKL algorithms utilize hand-crafted features such as Scale-Invariant Feature Transform (SIFT), Histogram of oriented gradient, and Bag-of-words [9], [10]. Different from the previous approaches, our study newly suggests a method to apply the features pre-trained by deep learning in unsupervised manner to MKL. Moreover, we suggest that MKL is expected to be used as analyzing the contribution of features from each layer for given machine learning problem by their optimized kernel combination weights. In this paper, we first discuss the core of MKL and Deep Belief Network in Section 2, and continue to elucidate the method to apply MKL to trained hierarchical features in Section 3. Next, we present the experimental procedures and results for classification tasks.

2 Preliminaries

2.1 Multiple Kernel Learning

Support Vector Machine (SVM) [11] is a large margin classifier that solves following optimization problems given training dataset $\{(x_1, y_1), (x_2, y_2), ...(x_N, y_N)\}$ where x_i is input vector $x \in \mathbb{R}^D$ and y_i is label $y_i \in \{-1, 1\}$ for $i = 1, 2, ..., N$

$$\begin{gathered}\min_{\boldsymbol{u},b} \frac{1}{2}(||\boldsymbol{u}||)^2 \\ \text{subject to } \; y_i(\boldsymbol{u}^T\phi(x_i) + b) \geq 1\end{gathered} \tag{1}$$

where $\boldsymbol{u}$ is weight of discriminant function $\boldsymbol{y} = \boldsymbol{u}^T\phi(\boldsymbol{x}) + b$, and b is bias.

The optimization problem can be represented by dual form with Lagrangian multipliers as shown below.

$$\begin{gathered}\max_{\boldsymbol{\alpha}} L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j\alpha_k y_j y_k(\phi(x_j^T)\phi(x_k)) \\ \text{subject to } \boldsymbol{\alpha} \geq 0, \sum_{i=1}^{N} \alpha_i y_i = 0\end{gathered} \tag{2}$$

The inner products of input data ($\phi(x_j^T)\phi(x_k)$) can be substituted with a kernel function $K(\phi(x_j), \phi(x_k))$. Also, multiple kernels are possible to be applied to SVM instead of single kernel. The combination of given subset of multiple kernels is mainly represented as the linear combination of kernels with the kernel combination parameter $\boldsymbol{\theta}$.

$$\begin{aligned} K(x_j, x_k) = \theta_1 K_1(\phi(x_j), \phi(x_k)) + \theta_2 K_2(\phi(x_j), \phi(x_k)) + ... \\ + \theta_M K_M(\phi(x_j), \phi(x_k)) \end{aligned} \tag{3}$$

where M corresponds to the number of kernels.

Multiple Kernel Learning (MKL) is proposed to optimize the combinations of kernels via finding the suitable $\boldsymbol{\theta}$ and simultaneously update the discriminative weights $\boldsymbol{u}$. As a result, MKL achieves the optimal combination of various kernels which measure the similarity of input in different manner or originate from diverse sources. In short, multiple features are able to be concatenated via MKL method. In addition, MKL can analyze the best kernel for the given machine learning problem.

2.2 Deep Belief Network

Deep Belief Network (DBN) is a multi-layer generative model of which building block is two-layer, undirected graphical model called Restricted Boltzmann machine (RBM). RBM consists of visible units $x_i \in \{0, 1\}$ and hidden units $h_j \in \{0, 1\}$. These units in one layer are fully-connected to units in the other layer, but there is no connection between units in the same layer. The energy function of RBM is defined as below.

$$E(\boldsymbol{x}, \boldsymbol{h}) = -\sum_{i=1}^{D} a_i x_i - \sum_{j=1}^{L} c_j h_j - \sum_{i=1}^{D}\sum_{j=1}^{L} h_j w_{i,j} x_i \tag{4}$$

where $\boldsymbol{w} \in \mathbb{R}^{D \times L}$ is the wieght vector between the visible layer and the hidden layer, $\boldsymbol{a}$ is the bias of the visible layer, and $\boldsymbol{c}$ is the bias of the hidden layer.

The joint probability of visible units and hidden units can be represented in the energy function.

$$P(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{Z} \exp\left(-E(\boldsymbol{x}, \boldsymbol{h})\right) \tag{5}$$

where Z is the partition function for normalization.

In order to train RBM model, the log-likelihood of the training data is maximized using gradient ascent learning by updating weights and biases of the model. The learning rule of weights $\boldsymbol{w}$ is proportional to the difference of the expectation based on training data and the expectation based on the model.

Since the model expectation at the above equation is intractable, Contrast Divergence [12] is proposed to approximate the model expectation. RBM employs Gibbs Sampling which alternatively samples one layer units given the other layer units and estimates based on the conditional probability.

3 Methods

Main idea of our approach is to combine different levels of abstraction of the given data, e.g. features, and consider them as multiple kernels. To implement our proposed method, we obtain hierarchical data representations with unsupervised deep learning. Then, we combine all the features from every level and train classifier via MKL.

The hierarchical feature representations are learned by DBN with greedily layer-wise training. We use DBN suggested by Hinton et al. [3] In addition, Convolutional DBN published by Honglak lee et al. [4] is also utilized to apply high dimension images. Convolutional DBN employs convolution and probabilistic max pooling method to accomplish the scalability.

In our methods, we obtain different levels of abstraction by estimating the activations of each level based on learned DBN weights for the given input data.

$$p(h_j^p = 1|\boldsymbol{x}^p) = \sigma(\sum_{i=1}^{N} w_{i,j}^p x_i^p + a_i^p) \tag{6}$$

where $\sigma()$ is the sigmoid function and p indicates pth layer.

In order to combine hierarchical feature representations learned and estimated under DBN model, we applied MKL in this study. There are many variations of MKL formulations, for instance, diverse forms of regularization terms and different optimization methods. We choose Ultra-Fast Online Multiple Kernel learning algorithms [13] that use the specific formation of regularization norm as below.

$$\Omega(\boldsymbol{u}) := \frac{\lambda}{2}||\boldsymbol{u}||^2_{2,\frac{2\log P}{2\log P-1}} + \alpha||\boldsymbol{u}||_{2,1} \tag{7}$$

where $\boldsymbol{u}$ is the weights of the kernel combinations, λ is a regulization coefficient, and P is the number of kernels equal to the number of layers in this paper. By selecting the regularization function, UFO MKL model can achieve optimal convergence rate only depending on the logarithm of the number of kernels.

In a nutshell, the combinations of hierarchical feature representations are optimized by UFO MKL algorithms. Each kernel function of MKL is replaced by each level of features from DBN.

4 Experiments

4.1 Databases

This study conducted experiments for handwritten digits database MNIST [14] and STL-10 database [15]. MNIST database has 10 classes of handwritten digits images, and the dimensions of each image are 28 by 28. STL-10 database contains 10 classes of objects images, including 100000 unlabeled images for unsupervised learning. The dimensions of each image in STL-10 are 96 by 96.

4.2 Handwritten Digits Classification

We train MNIST database with DBN with 3 layers. There is no fine-tuning of the whole level at the end of the layer-wise training. Each level of learned features is applied to the one of kernels in MKL. We use only 10000 images among 60000 training data images, because of memory limitation of current implementation. To compare the results, we also implement the experiments via SVM with concatenating features from all layers and only with features from the third layer. We uses LIBSVM [16] library in the experiments.

4.3 Objects Classification

The features for STL-10 database are learned by Convolutional DBN with 3 layers. The 96 by 96 images of STL-10 are resized to 32 by 32 images. The pre-training of the Convolutional DBN model is implemented by the provided unlabeled data of STL-10. With the pre-trained model, the features of training and test data for the classification are extracted. Each image is also preprocessed to have zero means and to be whitened with ZCA whitening. The hierarchical features are classified via MKL, where each kernel of MKL consists of each level of features. Besides, we conduct the classifications by SVM with concatenating features from all three layers and only with features from the third layer.

5 Results and Discussion

The classification results of MNIST handwritten digits are represented in Table 1. We compare the accuracy of the classification to prove the effectiveness of applying MKL to hierarchical features. The accuracy that the 3 levels of features are classified by MKL is 97.67 %, which is increased from the accuracy that the 3 layers of features are applied to SVM. The accuracy using the highest level of features is marked 96.84 % which is lower than both methods.

Table 1. The accuracy of the classification for MNIST database

Model	Accuracy (%)
3 layer features from DBN + MKL	97.67
3 layer features from DBN + SVM	96.96
Third layer features from DBN + SVM	96.84

The classification results for STL-10 database are descripted in Table 2. The overall results have the same tendency with the results of MNIST database. When MKL is utilized as classifier for 3 layers of features trained by DBN, the accuracy is improved than when SVM is used. When only the third layer features learned via Convolutional DBN (CDBN) without fine-tuning is classified by SVM, the accuracy is distinctively low.

Table 2. The accuracy of the classification for STL-10 database

Model	Accuracy (%)
3 layer features from CDBN + MKL	52.53
3 layer features from CDBN + SVM	51.10
Third layer features CDBN + SVM	18.81

The results validate that our approach with combined hierarchical features via MKL provides better performance than SVM or sole usage of the highest level of features. In short, MKL provides the way to find out the best combination of feature representations.

Table 3. The weights of kernels from the classification for MNIST database and STL-10 database

Layer	Weights for MNIST	Weights for STL-10
1	$0.3145 \times 10^{-4} \pm 0.55 \times 10^{-6}$	$0.0049 \pm 0.14 \times 10^{-4}$
2	0 ± 0	$0.0015 \pm 0.18 \times 10^{-4}$
3	$1.9238 \times 10^{-4} \pm 0.14 \times 10^{-6}$	0 ± 0

Moreover, the weight values of kernels in MKL mean which level of features contributes to solve the given machine learning problems. In MNIST and STL-10 classification tasks, the averaged weights of kernels from 10 trials of experiments are represented with standard deviations in Table 3. The weight corresponding to the second layer features converges to zero. Similarly, the weight of the third level of features converges to zero as well in the classification of STL-10 database. The low accuracy of the classification with the third layer features via SVM is explainable that the weight of the thirds layer features converges to zero. According to the results, the abstraction level of the second layer in MNIST features and the third layer in STL-10 features rarely contributes to represent the structure of the given data. Therefore, MKL is expected to be employed in analyzing the useful levels of features or appropriate levels of abstraction for the given machine learning tasks.

6 Conclusion

We suggest MKL with hierarchical feature representations method for classification task. By two classification experiments, applying MKL to the multiple levels of features pre-trained by deep learning is the better way to find the optimal combinations of hierarchical features. Furthermore, MKL is applicable to find proper levels of features to significantly contribute for the specific machine problem.

Acknowledgments. This research was supported by the Brain Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Plannig. (NRF-20120005790, NRF-2013035057) This work was supported by the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning as Global Frontier Project (CISS-20110031866, CISS-2012054204) This work was supported by the Research Center Program of IBS(Institute for Basic Science) in Korea. (HQ1201).

References

1. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning (2009)
2. Le Roux, N., Bengio, Y.: Representational power of restricted boltzmann machines and deep belief networks. Neural Computation 20, 1631–1649 (2008)
3. Hinton, G.E., Salakhutdinov, R.: Reducing the Dimensionality of Data with Neaural Netwoks. Science 313, 504–507 (2006)
4. Lee, H., Grosse, R., Ranganath, R., Ng, A.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th International Conference on Machine Learning, ICML (2009)
5. Bach, F., Lanckriet, G., Jordan, M.: Multiple kernel learning, conic duality, and the SMO algorithm. In: International Conference on Machine Learning (2004)
6. Yang, J., Li, Y., Tian, Y.: Group-sensitive multiple kernel learning for object categorization. In: Computer Vision (2009)
7. Vedaldi, A., Gulshan, V.: Multiple kernels for object detection. In: IEEE 12th International Conference on Computer Vision (2009)
8. Kloft, M., Rückert, U., Bartlett, P.L.: A Unifying View of Multiple Kernel Learning. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part II. LNCS, vol. 6322, pp. 66–81. Springer, Heidelberg (2010)
9. Binder, A., Nakajima, S., Kloft, M., Müller, C.: Insights from Classifying Visual Concepts with Multiple Kernel Learning. PloS One 7 (2012)
10. Nakajima, S., Binder, A., Müller, C.: Multiple kernel learning for object classification. In: Proceedings of the 12th Workshop on Information-Based Induction Sciences (2009)
11. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 297, 273–297 (1995)
12. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation 14, 1771–1800 (2002)
13. Orabona, F., Jie, L.: Ultra-fast optimization algorithm for sparse multi kernel learning. In: Conference on Machine Learning (2011)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradien-Based Learning Applied to Document Recognition. Proceedings of the IEEE (1998)
15. Coates, A., Lee, H., Ng, A.: An analysis of single-layer networks in unsupervised feature learning. Ann Arbor (2010)
16. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 1–39 (2011)
17. Salakhutdinov, R., Mnih, A., Hinton, G.: Restricted Boltzmann machines for collaborative filtering. In: Proceedings of the 24th International Conference on Machine Learning, ICML, pp. 791–798 (2007)

18. Kloft, M., Laskov, P., Zien, A.: Efficient and Accurate Lp -Norm Multiple Kernel Learning. In: Advances in Neural Information Processing Systems, NIPS (2009)
19. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2211–2268 (2011)
20. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation 1554, 1527–1554 (2006)
21. Swersky, K., Chen, B., Marlin, B., de Freitas, N.: A tutorial on stochastic approximation algorithms for training Restricted Boltzmann Machines and Deep Belief Nets. In: 2010 Information Theory and Applications Workshop (ITA), pp. 1–10 (2010)
22. Salakhutdinov, R.: Learning deep generative models (2009)
23. MNIST database of handwritten digits, `http://yann.lecun.com/exdb/mnist/`
24. STL-10 dataset, `http://cs.stanford.edu/~acoates/stl10`
25. LIBSVM, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Effects of Large Constituent Size in Variable Neural Ensemble Classifier for Breast Mass Classification

Peter Mc Leod and Brijesh Verma

Central Queensland University
Bruce Highway, North Rockhampton QLD 4702
mcleod.ptr@gmail.com, b.verma@cqu.edu.au

Abstract. This paper proposes a novel ensemble technique for mass classification in digital mammograms by varying the number of hidden units to create diverse candidates. The effects of adding more networks to the ensemble are evaluated on a mammographic database and the results are presented. A classification accuracy of ninety nine percent is achieved.

Keywords: Ensemble classifiers, neural networks, digital mammography.

1 Introduction

Breast cancer has increased in prevalence. The aetiology is unknown and a cure does not seem likely [1]. Research has progressed in relation to treatment but this relies on an accurate diagnosis however 11-25% of cancers are missed [2]. Reasons include distortion of the breast, occlusions with surrounding tissue, low mammogram contrast and even talc on the breast. The rate of breast cancer is low with three to four malignancies in a thousand [3]. A high volume of mammograms means that skill levels, complacency and fatigue can impact on radiologists. An estimated 35% of biopsies are not required [4] resulting in stress to patients and increased load on the health system. Despite this digital mammography is the diagnostic tool of choice due to wide availability, low cost and its non-invasive nature. Mechanisms such as a second radiologist to rescreen mammograms have been shown to improve the classification rate and reduce misdiagnosis. The cost and volume of mammograms makes this ineffective. Mechanisms including Computer Assisted Diagnostic (CAD) systems to act as an adjunct to radiologists have been suggested however variable classification accuracy has been a problem. This has been researched for around 40 years and arguably neural networks have demonstrated good capabilities. Techniques used to improve this situation include the use of many classifiers in a voting arrangement (ensemble). This research aims to create an accurate ensemble classifier.

This paper is broken into several sections with section 2 covering the research background. Section 3 details the proposed methodology while section 4 details the results. Discussions and analysis are in section 5 while section 6 details our conclusions and future research.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 525–532, 2013.

2 Background

Costa, Campos and Barros [5] used efficient coding based on Independent Component Analysis (ICA) achieving an accuracy of 90.07% on 5090 anomalies from the Digital Database of Screening Mammography (DDSM). They developed a compact code based on a statistics pattern ensemble to reduce redundancy with minimal loss of information. The data is transformed by linear functions generating an estimate of independent components. They used 41 components performing better than Principal Component Analysis (87.28% with 39 principal components) and Gabor Filter (85.28%). Luo and Cheng [6] used a bagged Decision Tree (DT) to gain an accuracy of 83.4% on mass anomalies. They utilized a DT and Support Vector Machine (SVM) Sequential Minimal Optimization. Mass anomalies from the University of California at Irvine (UCI) were classified using feature selection techniques to reduce the BI-RADS® input features from five to four. Mass margin was the most important feature. Their ensemble was more effective than using a single classifier. Yoon [7] achieved an area under the ROC curve of 0.94315 Az on a DDSM mass dataset with a boosted SVM ensemble together with fivefold cross validation to select the most appropriate features. Verma et al. used a partitioning mechanism for training of a classifier with direct output weight calculation by least squares (modified gram-schmidt) resulted in the creation of a Soft Clustered Neural Network (SCNN) [8] with 94% classification accuracy on mass anomalies from the DDSM. This technique removed those clusters that did not contribute to a class assignment in order to create a better decision boundary. The least squares technique does not suffer from local minima. Techniques of identifying sub-populations (soft-clusters) for the benign and malignant patterns to reduce class variability and increase classification accuracy on a neural network have also been used. This approach was known as Soft Clustered Based Direct Learning (SCBDL) [9] and achieved a classification accuracy of 97.5% on a dataset from the DDSM. Another approach used a SVM classifier with a genetic algorithm to select the classifier features [3]. This research attempted to test a new feature selection technique on a DDSM dataset with an accuracy of 89% being achieved. Other researchers examined mechanisms to create ensemble classifier; determining that 3-5 different classifiers were optimal taking into account diversity and variability [10].

3 Proposed Methodology

Neural networks are interconnected processing systems where each connection responds to input and the resultant outputs from the interconnected units (neurons) are aggregated to form a decision. Neural networks are capable of reaching a decision by the weights that interconnect the layers of neurons in the network. Through training knowledge of how to reach a decision is built into the weights.

Researchers examined the issue of obtaining the best possible configuration for neural networks with the selection of the best number neurons of being an area that was not investigated fully as the performance improvement was low. Investigations utilized only a small number of neurons in the hidden layer [11]. Others noted that too high a number was associated with overtraining [12, 13]. Diversity (or disagreement) is a key concept for the creation of ensembles. Diversity is the concept that a classifier is right more often than not; however when compared to another classifier its decision boundary is sufficiently different that it does not misclassify the same patterns. Combining the results of diverse classifiers should yield a result better than any single classifier. The proposed technique creates diverse classifiers to build an ensemble, as depicted in Figure 1. A detailed discussion of the system follows.

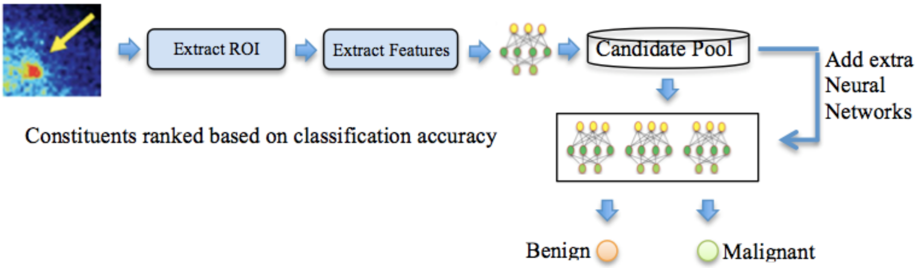

Extract ROI
Extract Features
Candidate Pool
Add extra Neural Networks
Constituents ranked based on classification accuracy
Benign
Malignant

Fig. 1. Proposed variable neuron based ensemble technique

3.1 Mammograms

The mammographic images for this research (100 malignant and 100 benign mass anomalies) are from the DDSM. This is one of the largest publicly available benchmark databases with 2600+ images. The anomalies are fully annotated with case information, cancer has been proven with biopsy and patients have been followed for a number of years to ensure that benign cases are indeed benign. Images are stored using a lossless compression algorithm, ensuring a high quality dataset.

3.2 Region of Interest

Mammographic images are large images to process and a diagnostic process is only concerned with making a diagnosis about a small area (anomaly). To conserve computational resources (memory and cpu capacity) only the Region Of Interest (ROI) (anomaly) is examined by extracting a boundary around the anomaly. The DDSM has a chain code for this process. Extracting the ROI does not attempt to classify an anomaly.

3.3 Feature Extraction

Once an anomaly has been extracted it is necessary to obtain the features that are used to form a decision as to whether it is malignant or benign. Breast masses are not easy to classify and no one feature can be used so multiple features are used. The features utilized in this research are based on the Breast Imaging Reporting and Data System

(BI-RADS®) as well as patient age and a subtlety value. BI-RADS® features have a positive predictive capacity for predicting mass malignancy [1, 14, 15]. The shape, density and mass margins are morphological features, which are utilized by radiologists. In some cases the pathology cannot be confirmed until histological samples are obtained and examined through biopsy. The difficult nature of performing a classification without a biopsy has been shown with the benign rate of biopsies being 65-90% [14]. Utilizing a feature set rather than a single feature increases classification accuracy however too many can reduce accuracy [16]. The features used in this research are patient age [17] (more aggressive tumors in younger patients), anomaly density (if the same density as surrounding tissue then hard to detect), shape (spiculated margins infer invasive tumors), margin (indistinct margins indicate harder to find and potentially more aggressive), subtlety (how hard is it to find) and assessment rank (a ranking of likely seriousness).

3.4 Network Training

A large number of neural networks are created by varying the number of neurons in a single hidden layer (from 2 to 1001) creating a large number of candidates for the ensemble. Changing the network parameters results in different weights between the layers, creating diverse classifiers. The candidates are created with the following parameters. Ten-fold cross validation is incorporated during training and testing. A Root Mean Square (RMS) error of 0.001 or (a maximum of 3000 iterations) is used for the stopping criteria. A learning rate of 0.05, momentum of 0.7 with six input neurons and two output neurons is used. Hyperbolic tangent sigmoid (tansig) is the transfer function between the layers with the system implemented in MATLAB™.

3.5 Ensemble Creation

The ensemble is created from the candidate pool with candidates ranked according to classification accuracy, which is the only inclusion mechanism. The first ensemble created is comprised of three neural networks. It is trained, tested and then another neural network is added with the process repeating to create a new ensemble of four neural networks (this is represented by the arrow in Figure 1.) This continues until an ensemble composed of 202 candidates is created. An upper bound of 202 is chosen to determine the effect of a large number of constituents (200 ensembles in total).

3.6 Classification and Fusion

Individual classifier results in the ensemble are fused together to form a classification using the majority vote algorithm, as it is one of the simplest but effective fusion mechanisms. In the event of a tie the smallest output value is chosen representing a malignant pattern. A false diagnosis for a malignant condition would be more severe than a false classification for a benign condition.

4 Experiments and Results

Experiments are conducted to create a candidate pool (one thousand) of back propagation neural network classifiers that had a different number of hidden units in the single hidden layer. It was hypothesized that this would be diverse enough to create an ensemble classifier with good accuracy.

Table 1. Performance of neural network on breast mass dataset (candidate classifiers)

Hidden Units	True Positive	False Negative	Accuracy
823	87	88	87.5
242	86	87	86.5
400	87	86	86.5
592	79	83	81.0
1000	82	78	80.0
78	69	81	75.0

Table 2. Performance of ensemble network on breast mass dataset

Constituents	Configuration	Accuracy (%)
3	823,242,400	95.0
4	823,242,400,24	92.5
10	823,242,400,24,262,302,404,657,5,15	97.5
100	823,242,400,24,262,302,404,657,5,15,32,268,281,292,309,494,550,31,43,50,75,158,165,183,209,224,349,355,356,398,416,426,436,443,466,473,622,639,659,661,678,749,903,904,925,38,59,68,79,116,146,168,175,204,218,223,232,233,235,243,246,254,277,297,304,305,325,350,352,366,388,395,417,427,444,459,471,493,500,537,546,556,583,612,664,682,739,753,842,866,870,887,930,957,999,14,30,37,95,103	98.5
127	823,242,400,24,262,302,404,657,5,15,32,268,281,292,309,494,550,31,43,50,75,158,165,183,209,224,349,355,356,398,416,426,436,443,466,473,622,639,659,661,678,749,903,904,925,38,59,68,79,116,146,168,175,204,218,223,232,233,235,243,246,254,277,297,304,305,325,350,352,366,388,395,417,427,444,459,471,493,500,537,546,556,583,612,664,682,739,753,842,866,870,887,930,957,999,14,30,37,95,103,104,138,140,166,171,174,187,188,202,212,221,252,259.282,288,312,340,367,368,384,391,421,438,470,510,551,573	99.0

Our literature review indicates that limited research into the creation of diverse networks by varying the number of hidden units in the hidden layer has been undertaken. The accuracy of the candidate networks ranged from 75% to 87.5%.

The candidate networks are ranked in descending order based on performance. The highest performers are selected for inclusion in the ensemble. Table 1 provides a summary of the classification accuracy achieved. Table 2 shows a subset of the accuracy achieved from the ensemble networks. Combining the best performing candidate networks created the ensemble.

5 Discussion

The results demonstrate that only a few candidates are needed to improve classification accuracy although this is variable in the early stages. To substantiate that an improvement in classification accuracy is achieved over the neural network an ANOVA analysis of variance is performed to see if the improvement is statistically significant (Table 3) using a 5% confidence level.

Table 3. ANOVA analysis summary

Groups	Count	Sum	Average	Variance
MLP	100	8485	84.85	0.335859
Ensemble	100	9815	98.15	0.063131

Table 4. ANOVA analysis details

	SS	*df*	*MS*	*F*	*P-value*	*F crit*
Between	8844.5	1	8844.5	44334.46	8.0331E-235	3.888853

In Table 4, the p-value is significantly below the confidence level confirming the improvement is statistically significant. The variance indicates the ensemble is more stable than the MLP network. Graphing accuracy of the ensemble against the number of classifiers shows a trend of higher accuracy as more classifiers are added. This levels off after around twenty classifiers (Figure 2). The highest classification accuracy of 99% is reached with 76 and 127 constituents. Stratification of the results is performed to determine the population variance as more classifiers are added.

Table 5. Ensemble variance, median and mode for ensemble groupings

No. Of Constituents	Variance	Median	Mode
3-12	3.10000	96.25	97.00
13-22	0.46944	97.25	97.50
53-62	0.19167	98.00	98.00
63-72	0.05556	98.00	97.50
163-172	0.10000	98.00	98.00
173-182	0.02500	98.00	98.00
183-192	0.04444	98.00	98.00
193-202	0.06944	97.75	98.00

A grouping of ten ensembles is chosen for each population in order to examine the changes of adding more classifiers. A subset of results is shown in Table 5. Variance tapers off as more classifiers are added (63-72 classifiers) then increases and tapers off again. In order to evaluate the performance of the proposed system it is necessary to compare its performance against that achieved by other researchers (Table 6).

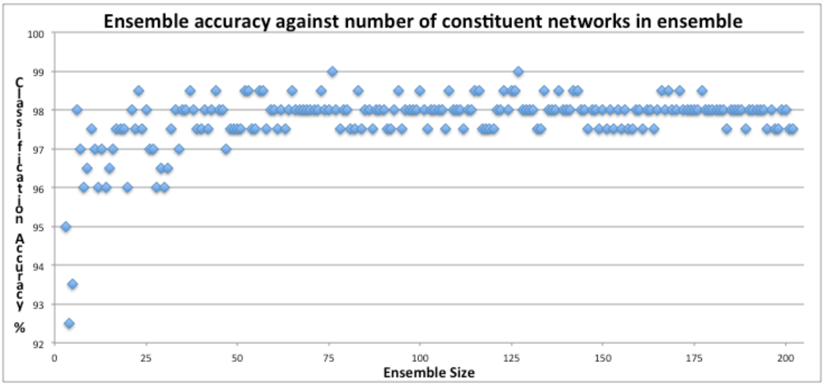


Fig. 2. Ensemble accuracy versus number of constituent classifiers

Table 6. Accuracy obtained by current research in comparison to proposed approach

Luo and Cheng [6]	Elfarra et al. [3]	Costa et al. [5]	Verma et al. [9]	Proposed
83.40%	89.00%	90.07%	97.5%	99.00%

6 Conclusions and Future Research

The variable neuronal ensemble has resulted in a high classification rate (99%) on the test dataset. This is high in comparison to other techniques. A disadvantage is that a high number of candidate networks are required to achieve high classification accuracy. It is noted that after a point adding more classifiers does not increase accuracy. This research uses a simplistic inclusion mechanism of accuracy. Our future research will use a multi-objective genetic algorithm with both diversity and accuracy.

References

1. Orel, S., Kay, N., Reynolds, C., Sullivan, D.: BI-RADS categorization as a predictor of malignancy. Radiology 211, 845–850 (1999)
2. Goergen, S., Evans, J., Cohen, G., Macmillan, J.: Characteristics of breast carcinomas missed by screening radiologists. Radiology 204, 131–135 (1997)
3. Elfarra, B.K., Abuhaiba, I.S.I.: New feature extraction method for mammogram computer aided diagnosis. International Journal of Signal Processing, Image Processing and Pattern Recognition 6, 13–36 (2013)

4. Isaac, L., Richard, L., Shalom, B., Yossi, S., Philippe, B., Fanny, S.: Computerized classification can reduce unnecessary biopsies in bi-rads category 4a lesions. In: Astley, S.M., Brady, M., Rose, C., Zwiggelaar, R. (eds.) IWDM 2006. LNCS, vol. 4046, pp. 76–83. Springer, Heidelberg (2006)
5. Costa, D., Campos, L., Allan, B.: Classification of breast tissue in mammograms using efficient coding. Biomedical Engineering OnLine 10 (2011)
6. Luo, S., Cheng, B.: Diagnosing breast masses in digital mammography using feature selection and ensemble methods. Journal of Medical Systems 36, 569–577 (2010)
7. Yoon, S., Kim, S.: AdaBoost-based multiple SVM-RFE for classification of mammograms in DDSM. In: IEEE International Conference on BioInformatics and Biomedicine (BIBMW 2008), Philadelphia, PA (2008)
8. Verma, B., McLeod, P., Klevansky, A.: A novel soft cluster neural network for the classification of suspicious areas in digital mammograms. Pattern Recognition 42, 1845–1852 (2009)
9. Verma, B., McLeod, P., Klevansky, A.: Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer. Expert Systems with Applications 37, 3344–3351 (2010)
10. West, D., Mangiameli, P., Rampal, R., West, V.: Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. European Journal of Operational Research 162, 532–551 (2005)
11. Partridge, D., Yates, W.: Engineering multiversion neural-net systems. Neural Computing 8, 869–893 (1996)
12. Hunter, D., Yu, H., Pukish, M.S.I., Kolbusz, J., Wiliamowski, B.M.: Selection of proper neural network sizes and architectures - A comparative study. IEEE Transaction on Industrial Informatics 8, 228–240 (2012)
13. Lawrence, S., Giles, C.: Overfitting and neural networks: conjugate gradient and backpropagation. In: International Joint Conference on Neural Networks, Como, Italy, pp. 114–119 (2000)
14. Vadivel, A., Surendiran, B.: A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories. Computers in Biology and Medicine 43, 259–267 (2013)
15. Mu, T., Nandi, A., Ranayyan, R.: Classification of breast masses using selected shape, edge-sharpness and texture features with linear and kernel-based classifiers. Journal of Digital Imaging 21, 153–169 (2008)
16. Kim, S., Yoon, S.: Mass lesions classification in digital mammography using optimal subset of BI-RADS and gray level features. In: 6th International Special Topic Conference on Information Technology Applications in Biomedicine, pp. 99–102 (2007)
17. Tabar, L., Fagerberg, G., Chen, H.-H., Duffy, S., Smart, C., Gad, A., et al.: Efficacy of breast cancer screening by age. New results swedish two-county trial. Cancer 75, 2507–2517 (1995)

Reduction of Ballistocardiogram Artifact Using EMD-AF

Ehtasham Javed, Ibrahima Faye, and Aamir Saeed Malik

Centre for Intelligent Signal & Imaging Research
Department of Electrical & Electronics Engineering
Universiti Teknologi PETRONAS, Perak, Malaysia
rajaehti1@gmail.com,
{Ibrahima_faye,aamir_saeed}@petronas.com.my

Abstract. Concurrent acquisition of functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) is widely used to monitor the neuronal activities of brain. However, this simultaneous recording suffers from complex artifacts. The Ballistocardiogram (BCG) artifact in specific, is as yet poorly assumed, appears to be more challenging and hinders to exploit the full strength of both modalities. In this paper, a hybrid method is implemented which combines Empirical Mode Decomposition (EMD) with Adaptive Filtering (AF) using notch filter to reduce the BCG artifact. Results of this study demonstrate that the proposed algorithm is generally useful and effective for the reduction of the BCG artifact.

Keywords: Simultaneous EEG-fMRI, Ballistocardiogram artifact, Empirical Mode Decomposition, Notch Filter.

1 Introduction

The quality of Electroencephalography (EEG) that can be attained from simultaneous acquisition with Functional Magnetic Resonance Imaging (fMRI) is still an on-going matter of investigation. The reduction of artifacts from EEG signal is essential to ensure full use of strengths of simultaneous acquisitions. The artifact which appeared to be more challenging for researchers to combine the strengths of EEG and fMRI is Ballistocardiogram (BCG) artifact which represents complex, non-linear and non-stationary characteristics [1].

The BCG artifact arises when the active circulatory system (endogenous contribution) interacts with the static magnetic field inside the MRI scanner (exogenous contribution) [2]. Ballistocardiogram artifact effect adds in the frequency range of the conventional EEG data, with the amplitude of 50μV (at 1.5 Tesla). Moreover, the BCG artifact is very similar to spikes of epilepsy [2]. It has duration of ~500 ms starting from Q wave of the Electrocardiogram (ECG). It has a rather complex and dynamic influence on the EEG signals.

Numerous researchers have proposed different methods for reduction of BCG artifacts. These include averaging [3], [4], adaptive filtering [5], Independent Component Analysis (ICA) [6], Principle Component Analysis (PCA) [1], [7], [8] and joint

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 533–540, 2013.

methods [4], [9]. The very first work on BCG removal was anticipated by Allen et al. [3]. They obtained a template window of artifact for each channel by calculating the average of the artifact per cardiac beat. The method was named as Average Artifact Subtraction (AAS). Bonmassar et al. in 2002 [5] used the adaptive Kalman filters and utilized motion sensors to measure the head movements to reduce BCG artifact.

The spatial methods like ICA and PCA have been suggested after their success in removal of ocular artifacts [7]. In the assessment of ICA, Optimal Basis Set (OBS) and OBS-ICA approaches performed by Debener et al. [8], it was observed that OBS and OBS-ICA provides better reduction of artifact.

Researchers also used other techniques, such as Kim et al. [9] who used a joint method based on wavelet denoising and filtering using adaptive recursive filters as post processing. Likewise, Adaptive Noise Cancellation following Optimal Basis Set is used by Niazy et al. [4]. In spite of several efforts to find an appropriate methodology for removing BCG artifact, quite a significant inconsistency exists between EEG-fMRI studies [8]. In this paper, a hybrid technique has been employed, in which the BCG artifact is reduced using Notch filter after applying Empirical Mode Decomposition (EMD).

The rest of the paper is organized as follows: Section 2 gives the theoretical background of the proposed method. The methodology is discussed in the section 3 and the results are presented in section 4 with validation. Section 5 consists of conclusion and future work.

2 Background

2.1 Empirical Mode Decomposition (EMD)

EMD was introduced to deal with both the nonlinear and non-stationary data. Unlike almost all existing methods; EMD is spontaneous, adaptive and does not require prior knowledge. The decomposed basis is based on, and results from the original signal [10]. The decomposed components are known as Intrinsic Mode Functions (IMFs). The IMFs have time-varying frequencies and amplitudes [10]. By definition the component will be called as IMF, if it satisfies two conditions which are as follows:

1. In complete dataset, the sum of total number of local maxima and local minima equals the number of zero crossings or their difference is at most one.
2. The local mean of lower and upper envelope is zero.

The IMFs can be extracted from original data using sifting process [10], described as;

1. Locate the local maxima and local minima of the original signal $X(t)$, and interpolate the extreme points via splines to obtain upper and lower envelopes.
2. Calculate the mean of two envelopes, m_1.
3. Obtain $\mathrm{h}_1 = \mathrm{X(t)} - \mathrm{m}_1$ and inspect the conditions for IMF.
4. If not, repeat the sifting process to obtain $\mathrm{h}_{1k} = \mathrm{h}_k - \mathrm{m}_{1k}$.
5. If h_{1k} constitutes an IMF, then designate it $\mathrm{c}_1 = \mathrm{h}_{1k}$.
6. Now we obtain the first residual r_1 via $\mathrm{r}_1 = \mathrm{X(t)} - \mathrm{c}_1$.

7. Treat r_1 as a new data set, and perform the sifting process to obtain c_2.
8. Continuing the sifting process we obtain

$$r_2 = r_1 - c_2, \dots, r_n = r_{n-1} - c_n.$$

9. Finally, the original signal is decomposed in terms of IMFs:

$$X(t) = \sum_{i=1}^{n} c_i + r_n \qquad (1)$$

2.2 Digital Notch Filter

Notch filters are more suitable for applications in which amplitude of a particular frequency needs to be reduced, but transmit remaining frequencies with no or minimal loss. Applications of notch filters are mainly in the field of communication and biomedical engineering. Removal of interferences from power line in the ECG recording system is an explicit example in the field of biomedical engineering [11].

Ideally, frequency response of digital notch filter is given by

$$\left|H(e^{j\Omega})\right| = \begin{cases} 0, & \text{for } \Omega = \Omega_0 \\ 1, & \text{for } \Omega \neq \Omega_0 \end{cases} \qquad (2)$$

Where $\Omega = 2\pi f/f_s$ is the normalized digital frequency, f_s is sampling frequency and Ω_0 denotes digital center notch frequency. The bandwidth of notch filter is the ratio of centre frequency f_0 and the quality factor Q, i.e.,

$$\Delta f = \frac{f_0}{Q} \qquad (3)$$

3 Proposed Methodology

A hybrid algorithm EMD-Adaptive Filtering (EMD-AF) is presented in this paper. In this algorithm, EMD is used to decompose the contaminated signal into components of different frequency and amplitude to differentiate the dynamic impact of BCG artifact on frequencies of EEG signal. AF is applied later to reduce the amplitude of BCG artifact from the decomposed components. The procedure of reducing BCG artifact is as follows:

1. EMD is applied to the contaminated signal to decompose it into different components of variable frequency and amplitude called as IMF.
2. Calculate the frequency spectrum of each IMF and compare their amplitudes to the prescribed threshold (maximum amplitude in original EEG signal, considered as acquired outside the scanner).
3. If the amplitude at certain frequency (range from 4-15 HZ i.e. typical BCG's frequency range [3]) is greater than the threshold value, apply notch filter to filter out the amplitude of that frequency in respective IMF.
4. Reconstruct the EEG signal without artifacts from the filtered IMFs using Equation (1).

The flow of the proposed algorithm is presented in Fig. 1.

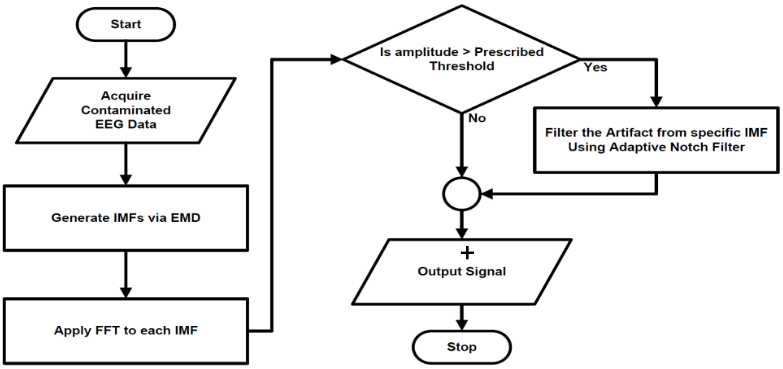

Start
Acquire Contaminated EEG Data
Generate IMFs via EMD
Apply FFT to each IMF
Is amplitude > Prescribed Threshold
Yes
No
Filter the Artifact from specific IMF Using Adaptive Notch Filter
+
Output Signal
Stop

Fig. 1. Flow chart of EMD-AF method

4 Results

The assessment of the proposed method in reduction of BCG artifact is the main purpose of this study. The proposed method is compared with the AAS. The parameters used to assess the performance are: correlation coefficient, signal to BCG artifact ratio (SBR) and power spectral density.

4.1 Simulated Data

In this simulated study, EEG data were taken from [12]. 8 channels were selected from 128 channels and are all from the right side of referenced electrode Cz. The selected data are used as the original (artifact-free) signal. The selected channels were grouped over 4 scalp regions as: Frontal (F), Fronto-Temporal (FT), Temporal (T) and Parietal (P) (two from each region). The reason of choosing electrodes from different regions is the fact that BCG amplitude varies significantly and there is inter-channel inconsistency in artifact morphology [2].

The BCG artifact was estimated by subtracting the reconstructed signal (obtained using default setting of BCG artifact removal toolbox in net-station) from the EEG recorded data inside the fMRI scanner at 3T of the same selected channels.

Four EEG signals (one from each above mentioned scalp regions) with estimated BCG artifact in respective channels are shown in Fig 2. Contaminated EEG data is created by mixing the eye-closed EEG data and estimated BCG artifact data. The proposed algorithm and AAS are applied on the (8 selected channels) contaminated data. The results are compared and validated in the next section.

4.2 Validation

First, the proposed algorithm and AAS are validated using correlation coefficient, which gives similarity between the reconstructed signal and the reference EEG

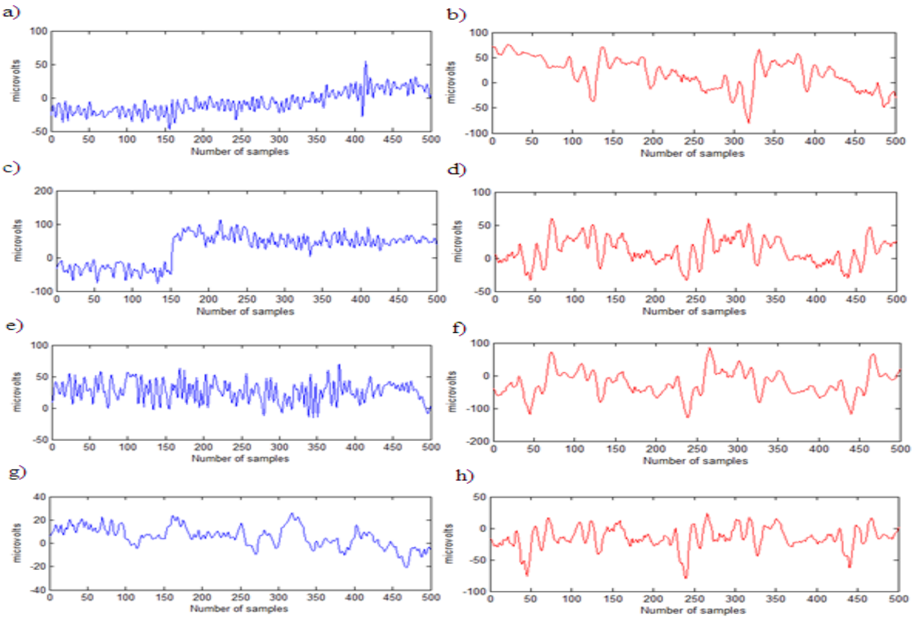


Fig. 2. Examples of simulated data used form four scalp region: eye-closed EEG signal from (a) Frontal, (c) Fronto-Temporal, (e) Temporal, (g) Parietal region and estimated BCG artifact in (b), (d), (f) and (h) respectively

waveform. Another parameter used to measure the degree of removal of BCG artifact is the Signal to BCG artifact ratio (SBR), calculated before and after the implementation of the both methods using:

$$\mathrm{SBR}_{\mathrm{contaminated}} = 10 * \log\left[\frac{\mathrm{RMS(o)}}{\mathrm{RMS(c)}}\right] \tag{4}$$

and

$$\mathrm{SBR}_{\mathrm{EMD-AF\ or\ AAS}} = 10 * \log\left[\frac{\mathrm{RMS(o)}}{\mathrm{RMS(r)}}\right] \tag{5}$$

Where ‘o’ is the original (artifact-free) EEG signal, ‘c’ is the contaminated signal and ‘r’ is the residual of BCG artifact, which remained in signal after applying the proposed method. r is calculated by $\mathrm{r = o - rs}$ and $'\mathrm{rs}'$ is the reconstructed signal.

Besides the SBR, we also calculated the relative root mean square error (RRMSE), with reference to SBR present in the signal after the reconstruction, using the formula:

$$\mathrm{RRMSE} = 10 * \log\left[\frac{\mathrm{RMS(r)}}{\mathrm{RMS(o)}}\right] \tag{6}$$

The region-wise average correlation coefficients using EMD-AF and AAS are presented in Table 1. The similarity index shows that the proposed method reduced most of the BCG artifact as compared to AAS from all regions. While comparison within the region shows that reconstructed signal has relatively high similarity in all regions except Parietal because the amplitude of BCG artifact is closer to the EEG, which can be seen from Fig 2 (g) and (h). Fig 3(a) shows that $\mathrm{SBR}_{\mathrm{EMD-AF}}$ has high ratio compared to $\mathrm{SBR}_{\mathrm{AAS}}$ and $\mathrm{SBR}_{\mathrm{contaminated}}$, which means EMD-AF reduced the artifact from contaminated signal to certain extent and better than AAS. The method works well on high SBR relative to the low SBR, which can be concluded from Fig 3(b), as RRMSE is very low at high SBR. Moreover, AAS does not show any consistency and has higher RRMSE than EMD-AF for all SBRs.

Power Spectral Density (PSD) of original signal, contaminated signal and reconstructed signals are compared to further evaluate the proposed method. Fig 4 shows that PSD of the reconstructed signal using EMD-AF is closer to the PSD of the original signal compared to the PSD of the reconstructed signal using AAS and contaminated signal. The difference in PSD of the original and the reconstructed signals shows that there are still residues of BCG artifact in the reconstructed signal via EMD-AF, which can also be concluded from similarity indexes presented in Table 1.

Table 1. Correlation coefficient over four scalp regions

Region	Channel	Correlation Coefficient		Mean	
		EMD-AF	AAS	EMD-AF	AAS
Frontal	1	0.6190	0.2310	0.69905	0.2388
	9	0.7791	0.2465		
Fronto-Temporal	115	0.6082	0.1881	0.7105	0.3523
	110	0.8128	0.5165		
Temporal	108	0.7416	0.2540	0.6064	0.1866
	96	0.4712	0.1192		
Parietal	98	0.4736	0.1553	0.4780	0.1970
	92	0.4825	0.2386		

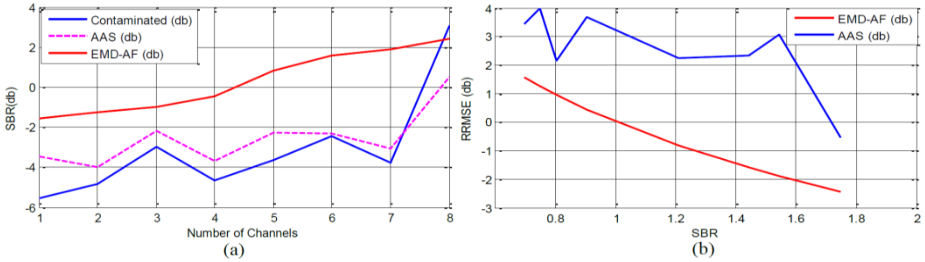


Fig. 3. (a) Signal to BCG artifact ratio (SBR) of contaminated signal (blue) and reconstructed signal (dotted pink) and (red) using AAS and EMD-AF of 8 channels respectively. (b) Relative Root Mean Square Error (RRMSE) after applying EMD-AF (red) and AAS (Blue) at different SBRs.

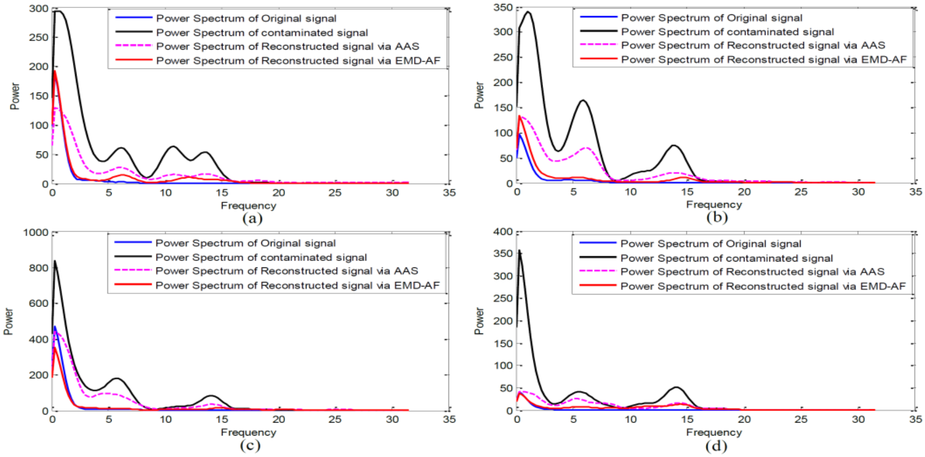


Fig. 4. Power Spectral density of original (blue), reconstructed (red) and (dotted pink) via EMD-AF and AAS respectively and contaminated (black) signals of one channel from each region: (a) Frontal, (b) Fronto-Temporal, (c) Temporal and (d) Parietal.

The results show that the proposed method has better capability to remove the BCG artifact while preserving the original EEG activities as compared to AAS. It needs to be further investigated and improved to fully reduce the BCG artifact to assess the neurological activities.

5 Conclusion and Future Work

A hybrid method for reduction of BCG artifact from EEG data acquired concurrently with fMRI is presented in this research article. The results of simulated signals from four different scalp regions show that the method can efficiently remove the BCG artifact, though not equally from all regions. The proposed method showed a good performance in reduction of BCG artifact at high SBR. Furthermore, the method does not require prior information about the sources mixed in the contaminated signal. However, while implementing on real data, the selection of a threshold value will require an EEG data acquisition without scanner. This method still needs improvements for a better reduction of BCG artifact and will be compared with the latest methodologies in our future works. In addition, it can be used jointly with other Blind Source Separation (BSS) methods.

References

1. Debener, S., Mullinger, K.J., Niazy, R.K., Bowtell, R.W.: Properties of the ballistocardiogram artefact as revealed by EEG recordings at 1.5, 3 and 7 T static magnetic field strength. International Journal of Psychophysiology 67, 189–199 (2008)
2. Debener, S., Kranczioch, C., Gutberlet, I.: EEG Quality: Origin and Reduction of the EEG Cardiac-Related Artefact. In: Mulert, C., Lemieux, L. (eds.) EEG - fMRI: Physiological Basis, Technique, and Applications, pp. 135–147. Springer, Heidelberg (2010)

3. Allen, P.J., Polizzi, G., Krakow, K., Fish, D.R., Lemieux, L.: Identification of EEG events in the MR scanner: the problem of pulse artifact and a method for its subtraction. NeuroImage 8, 229–239 (1998)
4. Niazy, R.K., Beckmann, C.F., Iannetti, G.D., Brady, J.M.: Removal of FMRI environment artifacts from EEG data using optimal basis sets. NeuroImage 28, 720–737 (2005)
5. Bonmassar, G.: Motion and Ballistocardiogram Artifact Removal for Interleaved Recording of EEG and EPs during MRI. NeuroImage 16, 1127–1141 (2002)
6. Mantini, D., Perrucci, M.G., Cugini, S., Ferretti, A., Romani, G.L., Gratta, C.D.: Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis. NeuroImage 34, 598–607 (2007)
7. Bénar, C.G., Aghakhani, Y., Wang, Y., Izenberg, A., Al-Asmi, A., Dubeau, F., Gotman, J.: Quality of EEG in simultaneous EEG-fMRI for epilepsy. Clinical Neurophysiology 114, 569–580 (2003)
8. Debener, S., Strobel, A., Sorger, B., Peters, J., Kranczioch, C., Engel, A.K., Goebel, R.: Improved quality of auditory event-related potentials recorded simultaneously with 3-T fMRI: removal of the ballistocardiogram artefact. NeuroImage 34, 587–597 (2007)
9. Kim, K.H., Yoon, H.W., Park, H.W.: Improved ballistocardiac artifact removal from the electroencephalogram recorded in fMRI. Journal of Neuroscience Methods 135, 193–203 (2004)
10. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 454, 903–995 (1998)
11. Ferdjallah, M., Barr, R.E.: Adaptive Digital Notch Filter Design on the Unit Circle for the Removal of Powerline Noise from Biomedical Signals. IEEE Transaction on Biomedical Engineering 41, 529–536 (1994)
12. Amin, H.U., Malik, A.S., Badruddin, N., Chooi, W.T.: EEG Mean Power and Complexity Analysis during Complex Mental Task. In: International Conference on Complex Medical Engineering (2013)

Exploring the Power of Kernel in Feature Representation for Object Categorization

Weiqiang Ren[1], Yinan Yu[2], Junge Zhang[1], and Kaiqi Huang[1]

[1] Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition (NLPR), CASIA
{wqren,jgzhang,kqhuang}@nlpr.ia.ac.cn
[2] Baidu Inc.
yuyinan@baidu.com

Abstract. Learning robust and invariant feature representations is always a crucial task in visual recognition and analysis. Mean square error (MSE) has been used in many feature encoding methods as a feature reconstruction criterion. However, due to the non-Gaussian noises and non-linearity structures in natural images, second order statistics like MSE are usually not sufficient to capture these information from image data. In this paper, motivated by the information-theoretic learning framework and kernel machine learning, we adopt a similarity measure called correntropy in the auto-encoder model to tackle this problem. The proposed maximum correntropy auto-encoder (MCAE) learns more robust and discriminative representations than MSE based model by performing computation in an infinite dimensional kernel space. Moreover, we further exploit the power of kernel by learning a kernel embedding neural network which explicitly maps data from Euclidean space to an approximated kernel space. Experimental results on standard object categorization datasets show the effectiveness of kernel learning in feature representation for visual recognition task.

Keywords: Auto encoder, maximum correntropy, explicit kernel embedding, image classification, feature learning.

1 Introduction

Feature learning/encoding is an essential step in the image classification. There have been a large number of feature encoding methods proposed to obtain better feature descriptions. One of the most successful framework for visual object categorization is the bag-of-features (BoF) model. The basic idea of BoF-based feature encoding methods is that local feature is expressed as a weighted linear combination of a couple of pre-trained visual words. Feature learning by these methods relies on a reconstruction of the original feature based on the mean square error (MSE) criterion, such as Sparse coding [1] and ICA [2], etc. Another state-of-the-art object recognition architecture is the so-called deep learning models [3], to name a few, RBM [4], Auto-Encoder [5], DBN [4], CNN [6]. Deeply rooted in the biological visual system and mathematical theory, deep

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 541–548, 2013.

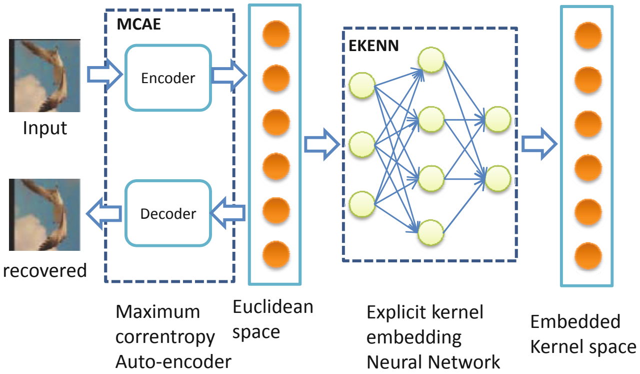


Fig. 1. Exploiting the power of kernel in feature learning

models mimic the cortex organization of human visual system by stacking one layer on top of another in a hierarchical way. Higher layer is able to learn more abstract concept than lower layer.

Analysis of natural images is hard as the distribution of natural images is highly non-uniform with non-Gaussian noise corruption and great variations. In order to handle the non-Gaussian noises and non-linear structures in natural images, we propose to train auto-encoder using the maximum correntropy criterion (MCC). Compared with the global mean square error (MSE) cost, correntropy [7] is a local similarity measure proposed in information theoretic learning (ITL). MCC usually obtains more robust feature description than MSE based methods in dealing with non-Gaussian noise data [8][9].

In order to learn more discriminative feature representation, we further exploit the power of kernel by explicit embedding of the Euclidean data into a kernel-induced feature space. The idea of explicit kernel mapping has drawn much attention recent years for large scale problems. Rahimi *el al.* [10] propose to find the explicit mapping by randomly sampling from the spectrum. Maji *el al.* [11] introduce an explicit kernel map for intersection kernel. Vedaldi *el al.* [12] further develop a general kernel mapping framework for homogeneous kernel. Unlike these explicit kernel mapping methods, our embedding method is capable of tackling more general kernels than a specific kernel or homogeneous kernels. Yu *el al.* [13] propose to train deep neural network by optimizing the hinge loss with kernel regularization, whereas we try to learn the explicit embedding guided by a kernel function.

The organization of this paper is as follows. In section 2, we introduce the preliminaries and propose the maximum correntropy auto-encoder (MCAE) model. In section 3 an explicit kernel embedding neural network is developed which further exploits the power of kernel for visual recognition. In section 4, we present the experimental results of the proposed model. Finally in section 5, we draw the conclusion and give the future lines of research.

2 Auto-Encoders for Robust Feature Learning

2.1 Auto-Encoders

Auto-Encoder is developed in the 1980s [14][15] as a special neural network architecture. Auto-encoder had been used as a powerful model for dimension reduction and feature extraction. In its basic format, an auto-encoder has two processing units, namely the encoder and the decoder. Usually there is a bottleneck layer, connecting the encoder and decoder, which is utilized as the learned feature for further processing. Many auto-encoder variants adopt the mean square error (MSE) with a different regularization term as the loss function, such as the sparse auto-encoders [16], contractive auto-encoders [17] and de-noising auto-encoders [18].

2.2 An Overview of Correntropy

Correntropy [7] is a localized similarity measure which is amenable for tackling non-Gaussian data with impulsive noises. For a given dataset $\{(x_i, y_i)|i = 1, 2, \cdots, N\}$, an estimation of the correntropy $H_{ce}(X, Y)$ between two random variables X and Y is defined as

$$H_{ce}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(x_i - y_i) \tag{1}$$

$$\kappa_\sigma(x_i - y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x_i-y_i)^2}{2\sigma^2}} \tag{2}$$

where $\kappa_\sigma(x_i - y_i)$ is the Gaussian kernel function.

2.3 Maximum Correntropy Auto-Encoder

Herein, we introduce the proposed unsupervised feature learning model called Maximum Correntropy Auto-Encoder (MCAE). Rather than using normal mean square error (MSE) or cross entropy as the reconstruction cost in training auto-encoder, we adopt a different criteria referred to as Maximum Correntropy Criterion (MCC) [7]. The objective of our MCAE model is as follows:

$$\mathcal{J}_{ce}(W, W', b, b') = -\frac{1}{2N}\sum_{m=1}^{N} \kappa_\sigma(g(f(\mathbf{x}^{(m)})) - \mathbf{x}^{(m)}) + \lambda(\|\mathbf{W}\|_2^2 + \|\mathbf{W}'\|_2^2) \tag{3}$$

where $f(\cdot)$ is the encoder and $g(\cdot)$ is the decoder. W, b and W', b' is the weights and biases of encoder and decoder. Note that the Gaussian kernel $\kappa_\sigma(\cdot)$ performs element-wise computation on each individual feature of $\mathbf{x}$. λ is the weight decay regularization parameter preventing overfitting of the model.

The proposed MCAE is directly related to the Information Theoretic Learning (ITL) framework [19]. Correntropy involves more higher order moments than normal second order statistics like mean square error (MSE). A reasonable conclusion is that more structural information in the training data can be captured by optimizing with respect to correntropy criterion. Moreover, the computational cost is relative low compared to moment expansions.

3 Explicit Kernel Embedding Neural Networks

In the previous section we introduce the maximum correntropy auto-encoder which processes information in the high dimensional kernel space implicitly. The learned features are fed into a classifier for classification task. SVM is a sophisticated classifier for image classification and is endowed with the power of kernel machine naturally. Linear SVM usually works fast and performs well for linear-separable problem. Kernel SVM, with the power of non-linearity, generally produces better results than the linear one. One of the main problems of kernel SVM is that the computation of gram matrix is very expensive especially when the dataset size is very large.

To retain the benefits of both linear and kernel SVM, a natural way is to seek an explicit mapping Φ that preserves the most properties of the implicit kernel space. Motivated by the expressive power of neural network, in this paper, we propose to use neural network as the approximation model to learn a explicit mapping function Φ with an Explicit Kernel Embedding Neural Network (EKENN).

Given a N layers neural network, the first layer is the input layer and the last layer is the output layer. There are N_l nodes in the lth layer. For a pair of training examples $(\mathbf{x}, \mathbf{y}), \mathbf{x}, \mathbf{y} \in \mathbb{R}^{N_1}$, we can perform feed-forward pass for $\mathbf{x}, \mathbf{y}$ individually. The activations at layer l are $\mathbf{a}^{(l,\mathbf{x})}$ and $\mathbf{a}^{(l,\mathbf{y})}$, respectively. As we want to approximate the kernel mapping function for a given kernel $\mathbf{K}(\cdot,\cdot)$, the dot product of the outputs from the last layer for $\mathbf{x}$ and $\mathbf{y}$ should be close to $K(\mathbf{x}, \mathbf{y})$. A natural choice for the loss function is defined as follows:

$$J(\mathbf{x}, \mathbf{y}, \mathbf{W}, \mathbf{b}) = \frac{1}{2}((\mathbf{a}^{(N,\mathbf{x})})^T \mathbf{a}^{(N,\mathbf{y})} - \mathbf{K}(\mathbf{x}, \mathbf{y}))^2 \tag{4}$$

We can readily train this model using back propagation algorithm.

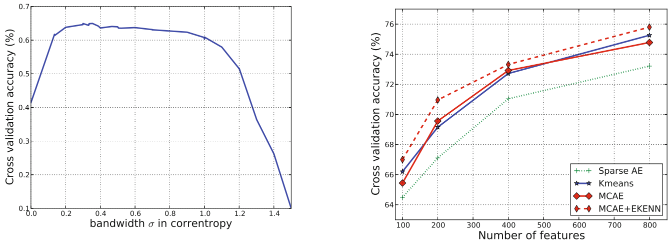
4 Experiments

In this section, we evaluate the proposed MCAE and EKENN on two standard image classification datasets, MNIST [6] and CIFAR-10 [20].

We choose L-BFGS for the optimization of auto-encoder. For the two-way back propagation in EKENN, we use SGD to learn the network as it is more convenient to handle pairs of training examples than L-BFGS. For EKENN, in the following experiments, we use 3-layers neural network to learn the non-linear mapping function induced by the χ^2 kernel.

4.1 Effect of Bandwidth σ

The bandwidth σ in our maximum correntropy auto-encoder model is an important parameter. In this experiment, we study the effect of bandwidth σ on CIFAR-10 datasets. The result is shown in Figure 2(a). From the experimental result, we can find that MCAE performs generally well when σ is in $[0.2, 0.6]$. In the following experiments, we set the value of σ by searching in this range for the best performance.



(a) Performance on CIFAR-10 with different σ.

(b) Performance on CIFAR-10 with different number of features.

Fig. 2. Experimental results on CIFAR-10

4.2 MNIST

MNIST [6] is a handwritten digit recognition dataset with 60000 training images and 10000 testing images. We train a MCAE with 1000 hidden units and report the classification error on testing set in Table 1. The results obtained before fine-tuning and after fine-tuning are presented separately. We also test the proposed EKENN by appending an extra processing block performing feature transformation on the fine-tuned features. The results show that even without fine-tuning MCAE still performs well. After fine-tuning, the test error drops to 1.23%. By using EKENN, we further improve the performance to 1.18%. The results we obtain are comparable to those from deep neural networks, as is shown in the last two rows of Table 1. For a better understanding of what MCAE learns, we draw the filters learned from MNIST in Fig. 3(a). It is clear that the MCAE model learns the strokes of the handwritten digits.

Table 1. Performance comparison on MNIST with 1000 features

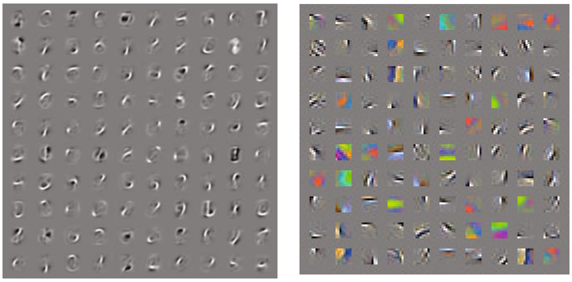
Algorithms	Test Error (%)
AE [17]	1.78
AE+wd [17]	1.68
DAE-b [17]	1.57
RBM [17]	1.30
CAE [17]	1.14
MCAE (no fine-tuning)	1.46
MCAE (fine-tuning)	1.23
MCAE (fine-tuning) + EKENN	1.18
Deep AE [5]	1.40
DBN [4]	1.20

4.3 CIFAR-10

CIFAR-10 [20] is composed of 60000 small 32 × 32 images in 10 classes. The whole dataset is divided into a training set with 50000 examples and a testing set with 10000 examples.

We follow the feature extraction pipeline in [21] on CIFAR-10. We randomly sample $100,000$ 8×8 images patches from the training set. The patches are first normalized to zero mean and unit standard derivation, followed by ZCA whitening and local contrast normalization (LCN). We use 2×2 division of the image and average pooling to formulate the final feature vector. When the number of features (i.e. the number of hidden units) is K, we actually extract a $4K$ dimensional feature vector as the representation for one image. We use LibLinear [22] to report the classification results and the regularization term C in SVM is chosen with cross-validation. In Fig 2(b) we report the performances of different feature learning algorithms under different number of features. As we can see, using larger number of features usually produces better performance. The result of MCAE is better than the MSE based sparse auto-encoder and comparable to kmeans. In Table 2 we list the results of MCAE and EKENN with 1600 features as well as results in the literature. Using MCAE and EKENN we get a classification accuracy of 78.8% on CIFAR-10. These results confirm the effectiveness of exploiting the power of kernels for feature representations.

The filters learned from CIFAR-10 is plotted in Fig. 3(b). Unlike those learned from MNIST, we find that the filters learned from the color CIFAR-10 images can be divided into two categories, the black-white patterns and the color patterns.



(a) MNIST (b) CIFAR-10

Fig. 3. Filters learned by MCAE

5 Conclusion

In this paper, we attempt to tackle the non-Gaussionity and non-linearity of image data and learn robust and discriminative feature representations. The proposed MCAE model adopts Maximum Correntropy Criterion (MCC) rather than mean square error (MSE) as the reconstruction cost. Besides, we propose a

Table 2. Results on CIFAR-10

Algorithms	Accuracy (%)
Sparse AE [21]	73.4
Improved LCC [23]	74.5
Kmeans [21] (1600 features)	77.9
MCAE (1600 features)	**77.6**
MCAE (1600 features) + EKENN	**78.8**

novel learning model that learns explicit kernel mapping in a data-driven fashion. Both the MCAE model and the explicit kernel embedding model work in a high dimensional kernel space induced by a kernel, which is regarded to be the main reason that it works better than other models that performs feature reconstruction in low dimensional Euclidean space. Experimental results on several widely used image classification datasets indicate the effectiveness of the proposed models. Our future work will be generalizing the MCAE model from MCC to more general kernel similarity cost, wherein we can put more prior information into the kernel similarity measure for better image modelling.

Acknowledgments. This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61175007), the National Key Technology R & D Program (Grant No. 2012BAH07B01).

References

1. Jianchao, Y., Kai, Y., Yihong, G., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1794–1801 (2009)
2. Le, Q.V., Karpenko, A., Ngiam, J., Ng, A.Y.: ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning. In: Advances in Neural Information Processing Systems (2011)
3. Bengio, Y.: Learning deep architectures for ai. Foundations and Trends® in Machine Learning 2(1), 1–127 (2009)
4. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)
5. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. Advances in Neural Information Processing Systems 19, 153 (2007)
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
7. Liu, W., Pokharel, P.P., Principe, J.C.: Correntropy: Properties and Applications in Non-Gaussian Signal Processing. IEEE Transactions on Signal Processing 55, 5286–5298 (2007)
8. Yuan, X.T., Hu, B.G.: Robust feature extraction via information theoretic learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pp. 1193–1200. ACM, New York (2009)

9. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1561–1576 (2011)
10. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. Advances in Neural Information Processing Systems 20, 1177–1184 (2007)
11. Maji, S., Berg, A.: Max-margin additive classifiers for detection. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 40–47. IEEE (2009)
12. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. Pattern Analysis and Machine Intellingence 34(3) (2011)
13. Yu, K., Xu, W., Gong, Y.: Deep learning with kernel regularization for visual recognition. In: NIPS 2008, pp. 1889–1896 (2008)
14. Rumelhart, D., Hintont, G., Williams, R.: Learning representations by back-propagating errors. Nature 323(6088), 533–536 (1986)
15. Baldi, P., Hornik, K.: Neural networks and principal component analysis: Learning from examples without local minima. Neural Networks 2(1), 53–58 (1989)
16. Hyvärinen, A., Hurri, J., Hoyer, P.: Natural Image Statistics: A Probabilistic Approach to Early Computational Vision, vol. 39. Springer (2009)
17. Rifai, S., Vincent, P., Muller, X., Glorot, X., Bengio, Y.: Contracting auto-encoders: Explicit invariance during feature extraction. In: Proceedings of the Twenty-Eight International Conference on Machine Learning (ICML 2011) (June 2011)
18. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103. ACM (2008)
19. Principe, J., Xu, D.: Information-theoretic learning using reny's quadratic entropy. In: First International Workshop on Independent Component Analysis (ICA 1999), pp. 407–412. Citeseer (1999)
20. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009)
21. Coates, A., Lee, H., Ng, A.: An analysis of single-layer networks in unsupervised feature learning. In: AISTATS 14, vol. 1001, p. 48109 (2011)
22. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
23. Yu, K., Zhang, T.: Improved local coordinate coding using local tangents (June 2010)

A Novel Image Quality Index for Image Quality Assessment

Sheikh Md. Rabiul Islam, Xu Huang, and Kim Le

Faculty of Education, Science, Technology and Mathematics
University of Canberra, Australia
{Sheikh.Islam,Xu.Huang,Kim.Le}@canberra.edu.au

Abstract. Image quality assessment (IQA) is provided as computational models to measure the quality of images in perceptually consistent manner. In this paper, a novel image quality index with highlighting shape of histogram of the image targeted is introduced to assess image qualities. The index will be used in place of existing traditional Universal Image Quality Index (UIQI) "in one go". It offers extra information about the distortion between an original image and a distorted image in comparisons with UIQI. The proposed index is designed based on modelling image distortion combinations of four major factors: loss of correlation, luminance distortion, contrast distortion and shape distortion. This index is easy to calculate and applicable in various image processing applications. Experimental results show that the proposed image quality index plays a significantly role in the quality evaluation on the open source "Wireless Imaging Quality (WIQ) database".

1 Introduction

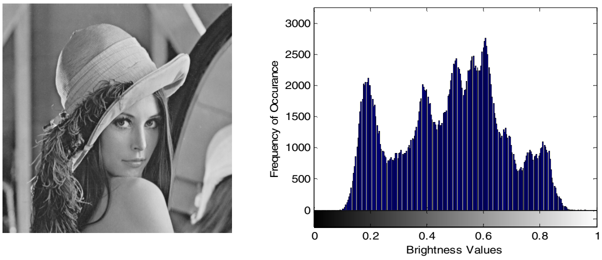
Image quality assessment (IQA) is an important issue for numerous image processing applications. Subjective IQA's performed by humans directly as it can give the most accurate quality assessment scores. Human eyes are the ultimate receivers in most image processing environment. However, the subjective evaluation methods are not only expensive and inconvenient but also very different to be integrated into computations. Therefore, it is desirable to develop objective methods which can automatically assess the quality of image with subjective results.

Over the past several decades research on this front has given rise to a variety of computation methods of image quality assessment. IQA may be classified three different types namely, full-reference (FR) (where the reference image is fully accessible when evaluating the distorted image), reduced-reference (RR) (where only partial information about the reference image is available), and non-reference (NR) (where there is no access to the reference image) [1].

The quality index proposed by Wang-Bovik [2] has been proven very efficient on image distortion evaluation. It considers three factors: loss of correlation, luminance distortion and contrast distortion, which are crucial in image quality measurement. Besides these three factors many studies show that in human visual system (HVS),

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 549–556, 2013.

information of image histogram plays a very important role, when human subjectively judges the quality of an image. In this paper, a new image quality index with highlighting shape of histogram will be introduced to assess image qualities. Histogram is a technique commonly used for image contrast enhancement. Histogram modelling techniques provide sophisticated methods of modifying the dynamic range and contrast of an image by altering each individual pixel. The image intensity histogram expresses a desired shape. In statistics, a histogram is a graphical representation showing a visual impression of the data distribution. It is used to show the frequency distribution of measurements. The total area of the histogram is equal to the number of data. The axis is generally specified as continuous, non-overlapping intervals of brightness values. The intervals must be adjacent and are chosen to be of the same size. A graphical representation of image histogram (Fig.1) displays the number of pixels for each brightness value in a digital image. Today, image histograms are presented on many modern digital cameras. Users can easily use them as an aid to show the distribution of tones captured. It is also shown that the image detail has been lost to blown-out highlights or blacked-out shadows.



(a) (b)

Fig. 1. Original test image Lena (a) from WIQ database (b) and its histogram

We integrate the shape of histogram into the Universal Image Quality Index metric. The index is the fourth factor added to existing Universal Image Quality Index (UIQI) to measure the distortion between original images and distorted images. Hence this new image quality index is a combination of four factors. The UIQI index approach does not depend on the image being tested and the viewing conditions of the individual observers. The targeted image is normally a distorted image with reasonable high resolution. We will consider a large set of images and determine a quality measurement for each of them. Some statistical methods or quality indexes are used to make an overall quality assessment via the proposed new image quality index. In this paper the performance evaluation of the proposed index will be tested on open source "Wireless Imaging Quality (WIQ) database" [3].The discussion of Full-reference (FR) methods.

The proposed image quality index will be compared with other objective methods. The image quality assessment has focused on the use of computational models of the human visual system [4]. Most human vision system (HVS)-based assessment methods transform the original and distorted images into a "perceptual representation" that takes into account near-threshold psychophysical properties. Wang et al. [5] and [6] measure structure based on a spatially localized measure of correlation in pixel values structural similarity (SSIM) and in wavelet coefficients MS-SSIM. Yalman proposed to use histogram based image quality index (HQI) in place of traditional error summation methods [7]. The noise quality measure (NQM) [8] described as the structural similarity (SSIM) index is motivated by the need to capture the loss of structure in the image.

The rest of this paper is organized as follows: Section 2 describes the proposed image quality index; Section 3 presents the experimental results. The paper ends with a brief conclusion.

2 Proposed Image Quality Index

The quality index proposed by Wang-Bovik [2] has been proven very efficient on image distortion performance evaluation. It considers three factors which are crucial in image quality measurement. Instead of using traditional error summation methods, Wang and Bovik's method [2] was designed to model any image distortion with a combination of three factors. More specifically, given two pixel gray level real-value sequences $x = \{x_1, \dots\dots\dots x_n\}$ and $y = \{y_1, \dots\dots\dots y_n\}$. They are obtained by the following equations:

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ \sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2,\ \sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$$

Where $\bar{x}$ is the mean of x, $\bar{y}$ is the mean of y $\sigma_x^2$is variance of x, σ_y^2 is variance of y and σ_{xy} is the covariance of x, y

Then, we can compute a quality factor, Q:

$$Q = \frac{4\sigma_{xy}\bar{x}\bar{y}}{(\bar{x}^2+\bar{y}^2)(\sigma_x^2+\sigma_y^2)} \tag{1}$$

Q can be decomposed into three components as

$$Q = \frac{\sigma_{xy}}{\sigma_x\sigma_y}\cdot\frac{2\bar{x}\bar{y}}{(\bar{x}^2+\bar{y}^2)}\cdot\frac{2\sigma_x\sigma_y}{(\sigma_x^2+\sigma_y^2)} \tag{2}$$

In equation (2), the first component is the correlation coefficient between x and y, which measures the degree of linear *correlation* between x and y. The second component measures how close the mean *luminance* is between x and y. The third component measures how similar the contrasts of the images are as σ_x and σ_y can be viewed as estimation of the *contrasts* of x and y. Hence, we have three components for a quality factor, Q which can be rewrite as:

$$Q = correlation \cdot luminance \cdot contrast$$

The values of the three components are in the range of $[0, 1]$. Therefore, the quality metric is normalized between$[0, 1]$.

Besides these three factors, many studies show that in human visual system (HVS), histogram of image information plays a very important role, when human subjective judges the quality of an image. It works by redistributing the gray-levels of the input image by using its probability distribution function. Although this method preserves the brightness in the output image with a significant contrast enhancement, it may produce images which do not look as natural as the input ones. To take the advantages of known characteristics of human perception, we introduce the shape of histogram with the Universal Image Quality Index metric. This proposed image quality of index will be tested throughout standard database of images.

We use the statistical differences to develop a novel image quality index. To find out the shape of histogram from distorted image, we computed its *kurtosis and skewness.* Skewness, indicating a degree of asymmetry of a histogram, is given by the following equation:

$$K_{Skewness} = \frac{\sum_{i=1}^{n}(y-\bar{y})^3}{(n-1)s^3} \tag{3}$$

Kurtosis quantifies a degree of histogram *peakiness* and tail weight. That is, data sets with high kurtosis tend to have a distinct peak near the mean and have a heavy tail. Data sets with low kurtosis tend to have a flat top near the mean. It can be described by the following equation

$$K_{Kurtosis} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^4}{(n-1)s^4} \tag{4}$$

The formula for modified skewness is:

$$K_{Modify\ Skewness} = \frac{\sum_{i=1}^{n}|y_i-\bar{y}|^3}{|n-1|s^3} \tag{5}$$

where n is the number of pixels at image distortion value y_i, $\bar{y}$ is the mean value of image distortion, s is the standard deviation.

The proposed new image quality index Q can be expressed by the four components as below:

$$Q = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \cdot \frac{2\bar{x}\bar{y}}{(\bar{x}^2+\bar{y}^2)} \cdot \frac{2\sigma_x\sigma_y}{(\sigma_x^2+\sigma_y^2)} \cdot \frac{2K_xK_y}{K_x^2+K_y^2} \tag{6}$$

The 4th component in the above equation measures how similar the shape of histogram of the images is. As K_x and K_y can be viewed as estimation of the shape of x and y, the values of the four components is normalized so this is still in the range of $[0, 1]$. The value of K_x and K_y are computed into three different way using equation (3), (4) & (5). Therefore the Q can be expressed by the following four factors:

$$Q = correlation \cdot luminance \cdot contrast \cdot shape$$

The new quality index will be applied to local regions using a sliding window for objective image quality analysis. For example starting from the top-left corner of the image, a sliding window with the size of $B \times B$ is moving pixel by pixel horizontally and then vertically through all pixels of the image. We assume that at the position of (i, j) in the target image, the local quality index Q_{ij} can be computed as equation (7). Here, the row number and column number of the image are n and m, then the overall normalized quality index is:

$$Q = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} Q_{ij} \tag{7}$$

The overall performance of the proposed image quality index is based on shape of histogram with UIQI [2] can be further described in Fig.2. In order to show the efficiency of this image quality index, the open source "Wireless Imaging Quality (WIQ) database [3]" is used for testing, which will be discussed in the next section.

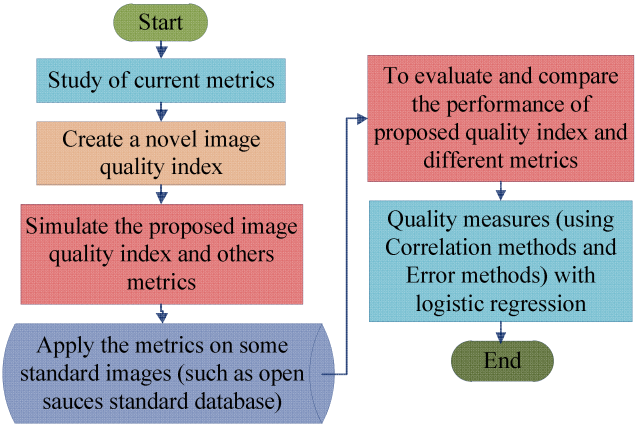

Start
Study of current metrics
Create a novel image quality index
Simulate the proposed image quality index and others metrics
Apply the metrics on some standard images (such as open sauces standard database)
To evaluate and compare the performance of proposed quality index and different metrics
Quality measures (using Correlation methods and Error methods) with logistic regression
End

Fig. 2. Flow chart of propose image quality index

3 Experimental Results

3.1 Image Database

We tested our proposed novel image quality index on the Wireless Imaging Quality (WIQ) assessment database [3]. This database contains seven widely adopted gray level reference images *(*Barbara, Elaine, Goldhill, Lena, Mandrill, Pepper, Tiffany) of dimensions 512 × 512 pixels. To address the problem of quality assessment for image communication they created a set of test images using a simulation model of a wireless link. This database has two subjective image quality tests. The first one is at the Western Australian Telecommunications Research Institute (WATRI) in Perth, Australia and the second one is at the Blekinge Institute of Technology (BTH) in Ronneby, Sweden. In each test, 30 non-expert viewers were presented with 40 test images. The artefacts in the test images were beyond what can usually be observed in purely source encoded images. In particularly, all the images are distorted by different types of blocking, blur, ringing, block intensity shifts, and high frequency noise. The database also contains the subjective evaluation results for each image, which is obtained by psychometric tests.

3.2 Methodology

To verify the validity and usefulness of the proposed image quality index, the experiments were performed with the following procedures:

Step 1: At first we obtained objective scores from the IQA algorithms. The evaluation was done using mean opinion score (MOS)/ difference mean opinion score (DMOS) after nonlinear regression using a five-parameter logistic function (a logistic function with an added linear term, constrained to be monotonic) as defined [9]:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1+e^{\beta_2 (x-\beta_3)}} \right) + \beta_4 x + \beta_5 \tag{8}$$

The parameters $\beta_1, \beta_2, \beta_3, \beta_4$ *and* β_5 are chosen to minimize the mean square error (MSE) values between predicted values (i.e. MOS/DMOS values) and the corresponding set of transformed predicted value $f(x)$. The minimization was conducted under the constraint that $f(x)$ had to be a monotonic function of over the range of predicted values. This nonlinearity was applied to the MOS or its logarithm, which ever gave a better fit for all data.

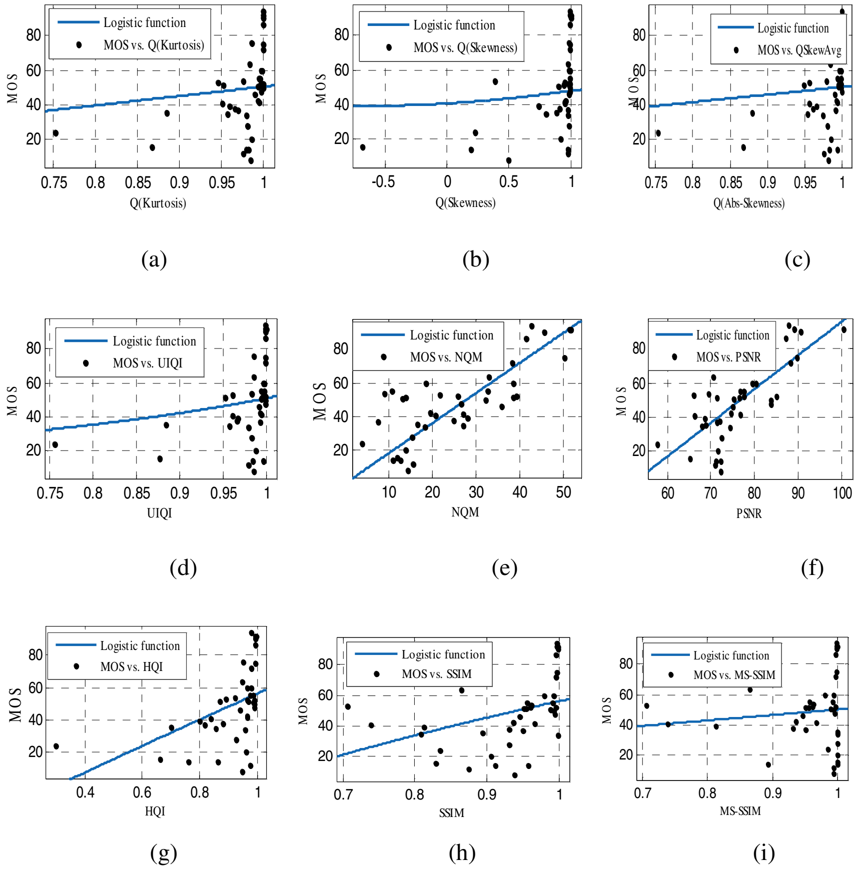
Step 2: According to the recommendations from Video Quality Experts Group (VQEG) [9], the performance of an image quality assessment can be quantitatively evaluated with respect to its ability to predict subjective quality rating in the following three aspects: (1) prediction accuracy, (2) prediction monotonicity, and (3) predication consistency. The first two metrics namely, Spearman rank-order correlation coefficient (SROCC) and Kendall rank-order correlation coefficient (KROCC), need to be measured with the prediction monotonicity of an IQA metric. These two metrics are based only on the rank of the data points with ignoring the relative distances among those points. The third metric is the Pearson linear correlation coefficients (PLCC) between the mean opinion scores and the objective scores after nonlinear regression. SROCC, KROCC, PLCC are the correlation methods. The fourth & fifth metrics are error methods such as the root-mean-square error (RMSE), and mean absolute error (MAE). These two are evaluated between the MOS and objectives scores after nonlinear regression. PLCC, MAE and RMSE are adopted to evaluate prediction accuracy. A better objective IQA measure is expected to have higher SROCC, KROCC and PLCC values (Prediction accuracy close to 1) while a low RMSE value (close to 0).

3.3 Experimental Results and Discussion

The proposed image quality index metric is generally competitive with the other metrics in terms of prediction accuracy and prediction monotonicity on the WIQ database. Here, the six metrics, peak signal to noise ratio (PSNR), UIQI [2], structural similarly(SSIM) [5], MS-SSIM [6], HQI [7], NQM [8] were applied .The results of PSNR, SSIM , MS-SSIM, HQI , NQM, UIQI were computed using their default implementation. Table 1 shows the simulation results of our proposed image quality index in three different cases such as image quality index Q (Skewness) for shape of histogram using skewness equation (3), image quality index Q (Abs-Skewness) for shape of histogram using absolute skewness equation (5), and image quality index Q (Kurtosis) for shape of histogram using kurtosis equation (4). It can be seen that the proposed method performs quite well for a wide range of distortion types. It gives better prediction accuracy, better prediction monotonicity & higher

Table 1. Evaluation of IQA models on WIQ database

Model	**Correlation Methods**			**Error Methods**	
	SROCC	KROCC	PLCC	MAE	RMSE
Q(Kurtosis)	**0.7049**	**0.5122**	**0.4116**	**92.7340**	**0.6550**
Q(Skewness)	**0.7297**	**0.5373**	**0.4568**	**92.7339**	**0.6552**
Q(Abs-Skewness)	**0.6551**	**0.5373**	**0.4568**	**92.7340**	**0.6551**
UIQI	**0.6551**	**0.4797**	**0.3795**	**92.7338**	**0.6554**
MSSIM	0.5781	0.4406	0.4182	92.7361	0.6559
MS-SSIM	0.3983	0.3259	0.2056	92.7333	0.6553
PSNR	0.6322	0.4929	0.7384	64.0075	0.3999
HQI	0.6708	0.4852	0.4705	92.7529	0.6560
NQM	0.6708	0.5109	0.7546	57.4466	0.3471



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 3. Scatter Plots of subjective MOS scores versus scores obtained by model prediction on the WIQ database (a) Proposed Q(Kurtosis); (b) Proposed Q(Skewness) (c) Proposed Q(Absolute Skewness); (d) UIQI; (e) NQM; (f) (PSNR); (g) HQI; (h) SSIM; (i) MS-SSIM.

MAE values & lower RMS values than UIQI and others methods, which is the most widely used FR image quality metric in the image processing literature. Fig.3 shows the scatter plots of MOS versus the predicted score by nine IQA metrics which achieve good results on the WIQ database. All curves shown in Fig.3 are obtained by a nonlinear fitting according to the model equation (8). It is clear that our proposed image quality index is more consistent with the subjective measure than others refer to the Fig.3 (a), (b) & (c) as compare with Fig.3 (d) & other methods by a nonlinear fitting using equation(8). The reason for the performance improvement is that we incorporate the measuring shape of histogram into the UIQI.

4 Conclusion

This paper introduced a novel image quality index. Our experimental results indicate that the proposed image quality index outperforms the UIQI under different types of image distortions. It is computationally efficient in comparison with typical image quality assessment (IQA) methods. It's also applicable to various input modalities. It does perform so well without any HVS model employed. Experimental results indicate that the proposed image quality index outperforms the others IQA models.

References

1. Wang, Z., Li, Q.: Information Content Weighting for Perceptual Image Quality Assessment. IEEE Trans. Image Process. 20(5), 1185–1198 (2011)
2. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Process. Lett. 9(3), 81–84 (2002)
3. Engelke, U., Zepernick, H.-J., Kusuma, T.M.: Wireless Imaging Quality Database (2010), `http://www.bth.se/tek/rcg.nsf/pages/wiq-db`
4. Taylor, C.C., Pizlo, Z., Allebach, J.P., Bouman, C.A.: Image quality assessment with a Gabor pyramid model of the human visual system, pp. 58–69 (June 1997)
5. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004)
6. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1398–1402 (2003, 2004)
7. Yalman, Y.: A Histogram based Image Quality Index. Przegl Ą Elektrotechniczny Electr. Rev. NR 7a(R. 88), 126–129 (2012)
8. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on a degradation model. IEEE Trans. Image Process. 9(4), 636–650 (2000)
9. VQEG, VQEG final report of FR-TV phase II validation test (2003)

Experimental Analysis of Moments of Predictive Deviations as Ensemble Diversity Measures for Model Selection in Time Series Prediction

Shuichi Kurogi, Kohei Ono, and Takeshi Nishida

Kyushu Institute of Technology, Tobata, Kitakyushu, Fukuoka 804-8550, Japan
{kuro@,ono@kurolab2.,nishida@}cntl.kyutech.ac.jp
http://kurolab2.cntl.kyutech.ac.jp/

Abstract. This paper presents an experimental analysis of moments of predictive deviations as measures of ensemble diversity to estimate the performance of time series prediction for model selection. As an extension of the ambiguity decomposition of bagging ensemble, we decompose the fourth power of ensemble prediction error and examine the effect of the moments of predictive deviations of ensemble members to the ensemble prediction error. By means of numerical experiments, we analyze the results to show the properties and the effectiveness of the moments.

Keywords: Moments of predictive deviations, Ensemble diversity measures, Performance estimation, Model selection, Time series prediction.

1 Introduction

This paper presents an experimental analysis of moments of predictive deviations of ensemble members as measures of ensemble diversity to estimate the performance of time series prediction for model selection. Here, the diversity, representing the degree of disagreement involved in the ensemble, has been well analyzed and applied to ensemble learning algorithms. For example, in NC (negative-correlation) learning of training dataset with known target values, an appropriate tuning of the trade-off between the minimization of each prediction error and the maximization of the covariance of prediction errors within the ensemble is shown to give a better performance (see [1,2]). However, with unknown target values, the performance of predictions is hard to be estimated. In [3], we have examined the fourth power of ensemble prediction errors and shown the effect of the moments of predictive deviations to the fourth power prediction error. From this analysis, we have shown that a trial model selection method using the maximum absolute skew of predictive deviations shows a better performance than the holdout method in multistep time series prediction. However, the analysis and the experiments are not enough to utilize the moments for model selection.

In this paper, we conducted numerical experiments much more and examine the effect of the moments of predictive deviations for model selection in time series prediction. In the next section, we show the notation of bagging, followed by the introduction

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 557–565, 2013.

of error decomposition of bagging ensemble prediction error and the moments of predictive deviations, and then describes multistep prediction of time series. In **3**, we show the results of numerical experiments and the effectiveness of the present analysis.

2 Bagging, Diversity and Time Series Prediction

2.1 Bagging for Regression Problem

Let $D^n \triangleq \{(\boldsymbol{x}_i, y_i) | i \in I^n\}$ be a training data set, where $\boldsymbol{x}_i \triangleq (x_{i1}, x_{i2}, \cdots, x_{Ike})^T$ and y_i denote an input vector and the target value, respectively, and $I^n \triangleq \{1, 2, \cdots, n\}$. We suppose the relationship given by

$$y_i \triangleq r_i + e_i = r(\boldsymbol{x}_i) + e_i, \tag{1}$$

where $r_i \triangleq r(\boldsymbol{x}_i)$ is a nonlinear target function of $\boldsymbol{x}_i$, and e_i represents zero-mean noise with the variance σ_e^2.

We formulate the bagging (bootstrap aggregation) [4,5] as follows; let $D^{n\alpha^\sharp, j} = \{(\boldsymbol{x}_i, y_i) |\ i \in I^{n\alpha^\sharp, j})\}$ be the jth bag (multiset, or bootstrap sample set) involving $n\alpha$ elements, where the elements in $D^{n\alpha^\sharp, j}$ are resampled randomly with replacement from the training dataset D^n. Here, α (> 0) indicates the bag size ratio to the given dataset, and $j \in J^{\text{bag}} \triangleq \{1, 2, \cdots, b\}$. Here, note that $\alpha = 1$ is used in many applications (see [6,7]), but we use variable α for improving generalization performance (see [5] for the effectiveness and the validity). With multiple learning machines θ^j $(\in \Theta^{\text{bag}} \triangleq \{\theta^j | j \in J^{\text{bag}}\})$ which have learned $D^{n\alpha^\sharp, j}$, the bagging for estimating the target value $r_i = r(\boldsymbol{x}_i)$ is done by

$$\hat{y}_i^{\text{bag}} \triangleq \hat{y}^{\text{bag}}(\boldsymbol{x}_i) \triangleq \frac{1}{b} \sum_{j \in J^{\text{bag}}} \hat{y}_i^j \equiv \left\langle \hat{y}_i^j \right\rangle_{j \in J^{\text{bag}}} \tag{2}$$

where $\hat{y}_i^j \triangleq \hat{y}^j(\boldsymbol{x}_i)$ denotes the prediction by the jth machine θ^j. The angle brackets $\langle \cdot \rangle$ indicate the mean, and the subscript $j \in J^{\text{bag}}$ indicates the range of the mean. For simple expression, we sometimes use $\langle \cdot \rangle_j$ instead of $\langle \cdot \rangle_{j \in J^{\text{bag}}}$ in the following.

2.2 Error Decomposition and Moments of Predictive Deviations

To analyze the error of bagging ensemble, bias-variance decomposition and ambiguity decomposition have been examined [1]. We here show a slightly different formulation in order to deal with unknown target values as follows. First, let us decompose each prediction as $\hat{y}_i^j = r_i + \beta_i + \epsilon_i^j$, where $\beta_i = \langle \hat{y}_i^j \rangle_j - r_i$ represents the bias, and $\epsilon_i^j = \hat{y}_i^j - \beta_i$ the predictive deviation. Then, we have the mean square error of the predictions $\hat{y}_i^j$ for all bags to the training target value y_i as

$$\left\langle (\hat{y}_i^j - y_i)^2 \right\rangle_j = \left\langle (\beta_i + \epsilon_i^j - e_i)^2 \right\rangle_j = (\beta_i)^2 + \left\langle (\epsilon_i^j)^2 \right\rangle_j - 2\beta_i e_i + (e_i)^2, \tag{3}$$

and the square error of the bagging prediction to the true target value, which we sometimes call generalization error in the following, as

$$(\hat{y}_i^{\text{bag}} - r_i)^2 = (\beta_i)^2 = \left\langle (\hat{y}_i^j - y_i)^2 \right\rangle_j - \left\langle (\epsilon_i^j)^2 \right\rangle_j + 2\beta_i e_i - e_i^2. \tag{4}$$

Here, (3) corresponds to the bias-variance decomposition and (4) the ambiguity decomposition, where the variance term $\langle (\epsilon_i^j)^2 \rangle_j$ is called ambiguity as a measure of diversity. Differently from the decompositions shown in [1], the above decompositions show the effect of overfitting $\beta_i e_i$ which should be taken into account for reducing the generalization error although it is hard to be estimated. Intuitively from the ambiguity decomposition, larger variance is supposed to reduce the generalization error much more. Furthermore, since we can obtain only the variance term when predicting unknown y_i, we expect that the variance is useful to estimate the generalization error. However, the variance has no relationship with the generalization error in our experiments shown below, which is of course owing that the first term may cancel the effect of the variance and it is well known that the variance becomes bigger as the complexity of the learning model increases.

So, we decompose the fourth power of the error as follows:

$$(\hat{y}_i^{\text{bag}} - r_i)^4 = -\left\langle (\epsilon_i^j)^4 \right\rangle_j - 4(\hat{y}^{\text{bag}} - y_i)\left\langle (\epsilon_i^j)^3 \right\rangle_j - 6(\hat{y}^{\text{bag}} - y_i)^2 \left\langle (\epsilon_i^j)^2 \right\rangle_j + C \tag{5}$$

where C represents the sum of the terms which do not explicitly involve ϵ_i^j, and note that $\hat{y}^{\text{bag}} - y_i = \beta_i - e_i$ involves unknown y_i.Then, in order to reduce the right hand side for a constant C, both $\langle (\epsilon_i^j)^4 \rangle_j$ and $\langle (\epsilon_i^j)^2 \rangle_j$ should be larger while $|\langle (\epsilon_i^j)^3 \rangle_j|$ should be smaller and larger for the corresponding terms being negative and positive, respectively. Now, to evaluate the relationship to the generalization error without any dependency among these terms, we examine the following moments of predictive deviations, i.e. the skew S_i and the kurtosis K_i as well as the variance V_i,

$$V_i \triangleq \sigma_i^2 \triangleq \langle (\epsilon_i^j)^2 \rangle_j, \quad S_i \triangleq \frac{\langle (\epsilon_i^j)^3 \rangle_j}{\sigma_i^3}, \quad K_i \triangleq \frac{\langle (\epsilon_i^j)^4 \rangle_j}{\sigma_i^4}. \tag{6}$$

For all test data to be predicted, we use mean variance (MV) $\langle V_i \rangle_i$, mean absolute skew (MS) $\langle |S_i| \rangle_i$ and mean kurtosis (MK) $\langle K_i \rangle_i$, where we do not use $\langle S_i \rangle_i$ but $\langle |S_i| \rangle_i$ to eliminate the effect of unknown polarity of $(\hat{y}^{\text{bag}} - y_i)\langle (\epsilon_i^j)^3 \rangle_i$.

2.3 Time Series Prediction

The above analysis is utilized for estimating the performance of multistep prediction of time series formalized as follows. Let $y_{t:n} = y(t)y(t+1)\cdots y(t+n-1)$ denote a time series of real values $y(t)$ ($\in \mathbb{R}$) for a discrete time $t = 0, 1, 2, \cdots$. For a given time series $y_{t_{\text{g}}:n_{\text{g}}}$, we are supposed to predict succeeding time series $y_{t_{\text{p}}:n_{\text{p}}}$ for $t_{\text{p}} \geq t_{\text{g}} + n_{\text{g}}$. To solve the problem, we use $y_t = r(\boldsymbol{x}_t) + e_t$ in (1) with substituting $y_t := y(t)$ and $\boldsymbol{x}_t := (y(t-1), y(t-2), \cdots, y(t-k))^T$, where the embedding dimension k should be selected properly (see the theory of Chaotic time series [8] for details).

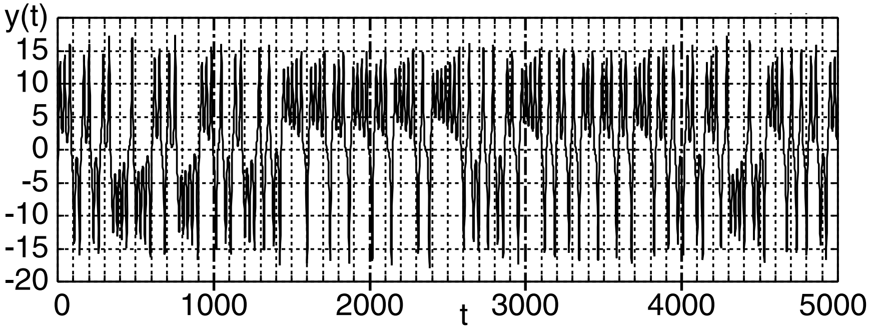


Fig. 1. Lorenz time series $y(t)$ for $t = 0, 1, 2, \cdots, 4999$

Then, the learning and the prediction can be formulated as a regression problem as described above, and we can execute multistep prediction $\hat{y}_t = \hat{y}^{\rm bag}(\hat{\boldsymbol{x}}_t)$ with $\hat{\boldsymbol{x}}_t = (x_{t1}, x_{t2}, \cdots, x_{tk})$ involving $x_{tj} = y_{t-j}$ $(t - j < t_{\rm g})$ and $x_{tj} = \hat{y}_{t-j}$ $(t - j \geq t_{\rm g})$ for $t = t_{\rm p}, t_{\rm p} + 1, \cdots$, successively.

3 Numerical Experiment and Analysis

3.1 Experimental Settings

As a chaotic time series, we employ Lorenz time series given by

$$\frac{dx}{dt_c} = -\sigma x + \sigma y, \quad \frac{dy}{dt_c} = -xz + rx - y, \quad \frac{dz}{dt_c} = xy - bz, \tag{7}$$

for $\sigma = 10$, $b = 8/3$, $r = 28$ (see [8]). Here, we use t_c for continuous time and t $(= 0, 1, 2, \cdots)$ for discrete time related by $t_c = tT$ with the sampling period T. We have generated 5,000 data points from the initial state $(x(0), y(0), z(0)) = (-8, 8, 27)$ with the sampling period $T = 25$ms via Runge-Kutta method with 128 bit precision of GMP (GNU multi-precesion library). We use $y(t)$ for the time series to be processed (see Fig. 1). Here, note that we have observed three time series generated with $T = 250$ms, 25ms and 2.5ms, respectively, and they are all the same until 20s and the latter two time series with $T = 25$ms and 2.5ms are the same until 30s, while the difference increases exponentially after then. Furthermore, with the precision less than 128 bit, the difference of the above time series increases after shorter duration of time. This result is supposed to be related to the property of chaotic time series that a short term prediction is possible but a long-term prediction is impossible owing to finite computational precision. From another point of view, the result indicates that the computational error by Runge-Kutta method decreases by reducing the sampling period, and $y(t)$ for each duration of time less than 1200 steps (=30s/25ms) in Fig. 1, or $y_{t_0:1200}$ for each initial time $t_0 = 0, 1, 2, \cdots$ with initial state $(x(t_0), y(t_0), z(t_0))$, is supposed to be almost correct, while cumulative computational error may increase exponentially after the duration.

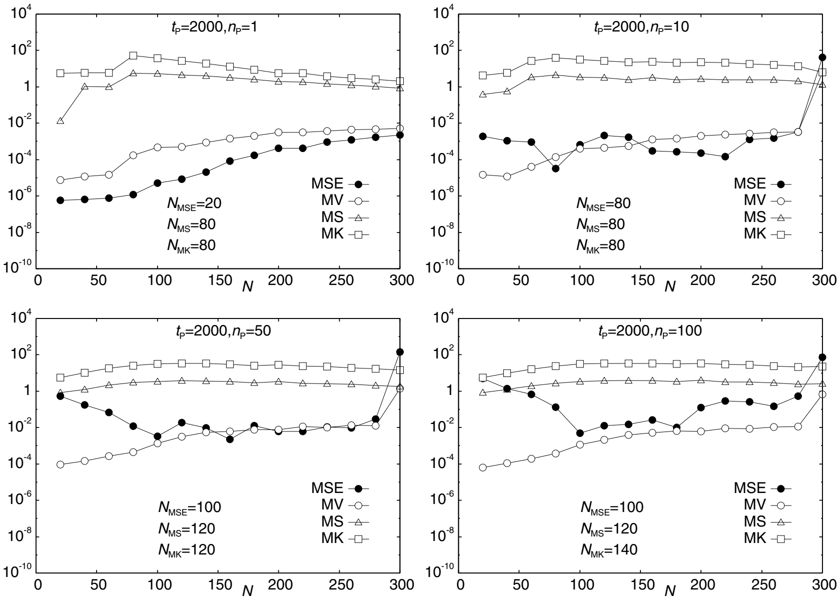


Fig. 2. Experimental results of MSE (mean square error), MV, MS, and MK vs. N for $t_{\mathrm{p}} = 2000$

We use $y_{0:2000}$ for training a bagging learning machine, and execute multistep prediction of $y_{t_{\mathrm{p}}:n_{\mathrm{p}}}$ with the initial input vector $\boldsymbol{x}_{t_{\mathrm{p}}} = (y(t_{\mathrm{p}} - 1), \cdots, y(t_{\mathrm{p}} - k))$ for prediction start time $t_{\mathrm{p}} = 2000 + 100i$ $(i = 0, 1, 2, \cdots, 29)$ and prediction horizon $n_{\mathrm{p}} = 1$, 10, 50, 100. Finally, we analyze the moments of predictive deviations of the bagging machines. Note that the relationship between training and predicting data of this experiment is different from usual one which employs $t_{\mathrm{p}} = t_{\mathrm{g}} + n_{\mathrm{g}}$. However, we can obtain several critical properties of the moments, as shown below.

For a learning machine, we use CAN2 (see **A.2** and [5] for details), where the model complexity is the number of units, N, or the number of piecewise linear regions for approximating a predictive function. We would like to solve the problem to select an appropriate N for a good prediction from $N \in \mathscr{N} = \{20, 40, \cdots, 300\}$. We use the embedding dimension $k = 8$ and the number of bags $b = 100$ because they have provided good prediction performance in several trial experiments.

3.2 Results and Analysis

For all 120 combinations of $t_{\mathrm{p}} = 2000 + 100i$ $(i = 0, 1, 2, \cdots, 29)$ and $n_{\mathrm{p}} = 1, 5, 10, 100$, we have executed the bagging prediction and obtained the mean square prediction error (MSE), and the moments of predictive deviations, i.e. mean variance (MV), mean absolute skew (MS) and mean kurtosis (MK) for all $N \in \mathscr{N} = \{20, 40, \cdots, 300\}$. Experimental result of the (mean) moments for $t_{\mathrm{p}} = 2000$ and $n_{\mathrm{p}} = 1$, 10, 50, 100 is shown in Fig. 2, where N_{MSE}, N_{MS} and N_{MK} denote N which achieves the minimum MSE, the maximum MS and the maximum MK, respectively, for all N and each t_{p} and n_{p}.

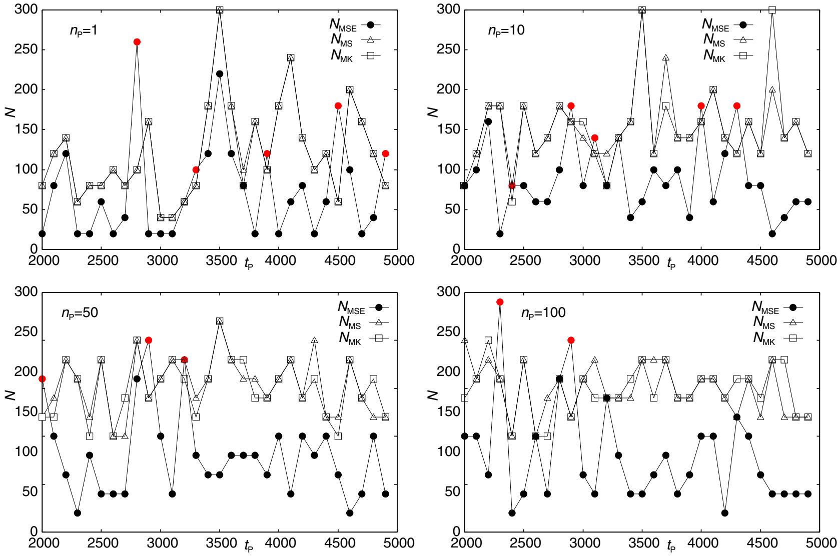


Fig. 3. Experimental results of N_{MSE}, N_{MS} and N_{MK} for $t_{\mathrm{p}} = 2000 + 100i$ $(i = 0, 1 \cdots, 29)$

By means of examining the results for all t_{p} and n_{p}, we have observed several properties. First, for the increase of N from 20 to 300, MS and MK increase to the global maximum and then decreases while MV only increases from the minimum to the maximum, where the increase and the decrease sometimes involve fluctuations but changes monotonically. This property is supposed to be obtained as follows. Basically, all moments of predictive deviations are supposed to increase with the increase of N because N is the number of piecewise linear regions providing the model complexity and then the predictive deviations may become larger with the increase of the complexity. When N is so big that the number of training data in each piecewise linear regions may not increase, all moments of predictive deviations are supposed to saturate. Since the skew S_i and the kurtosis K_i are normalized via using the value of variance $V_i = \sigma_i^2$ as given in (6), they may decrease when the variance, or the 2nd moment, decreases much more than the 3rd and the 4th moments. This may plausible because the minimization of the mean square prediction error, or the 2nd moment of the prediction error, for each N and training data is the objective of the learning of each CAN2, which is the same as usual learning machines.

Next, we have observed the property that N_{MSE} is smaller than N_{MS} and N_{MK} in almost all cases as shown in Fig. 3, where 12% cases (15 out of 120=30×4 cases) are the exceptions. In Fig. 3, we can see that N_{MSE}, or the best model, changes largely for the change of the start time t_{p} and the horizon n_{p} for prediction. From Fig. 3, we would like to evaluate the performance by means of using the half of N_{MS} and N_{MK} as a rough estimation of N_{MSE}. The result is shown in Fig. 4. Here, $\mathrm{MSE}_{\mathrm{min}}$, $\mathrm{MSE}_{\mathrm{MS}}$ and $\mathrm{MSE}_{\mathrm{MK}}$ are MSE of the predictions using N_{MSE}, $N = 0.5N_{\mathrm{MS}}$ and $N = 0.5N_{\mathrm{MK}}$, respectively. For comparison, $\mathrm{MSE}_{\mathrm{ref}}$ is obtained by applying $N = N_{\mathrm{MSE}}$ obtained for $y_{2000+100i):n_{\mathrm{p}}}$ to the prediction of $y_{2000+100(i+1):n_{\mathrm{p}}}$. This method is based on the

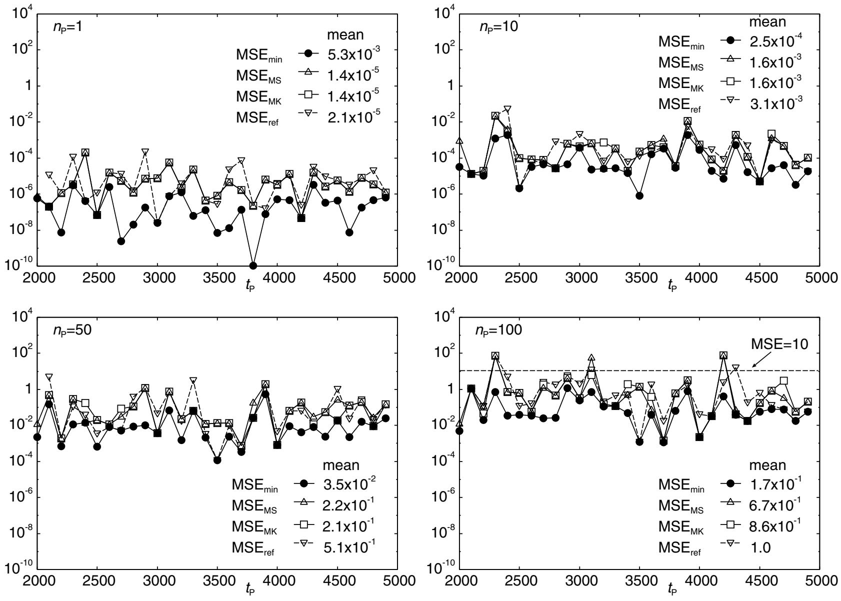


Fig. 4. Experimental results of MSE obtained by presented model selection methods

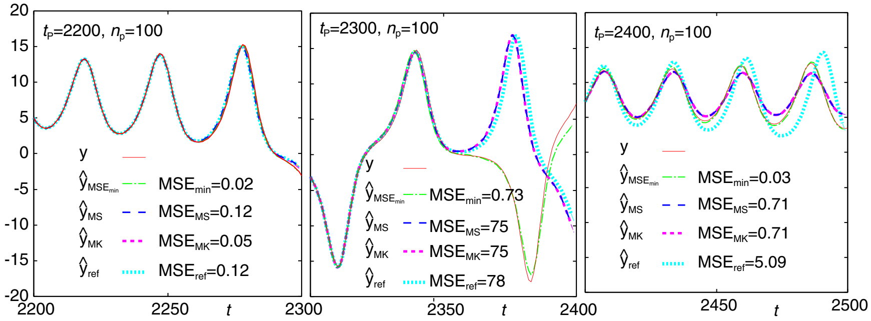


Fig. 5. Experimental results of predictions obtained by presented model selection methods

assumption of the continuity of the best model with respect to the increase of t_p, and this assumption is considered to be utilized by the holdout method (**A.1**) often used in time series model selection. Note that in Fig. 4 for $n_\mathrm{p} = 100$, we exclude the predictions using N which generates MSE ($\mathrm{MSE_{MS}}$, $\mathrm{MSE_{MK}}$ or $\mathrm{MSE_{ref}}$) bigger than 10 from the calculation of mean MSEs, because MSE bigger than 10 dominates the mean value. Precisely, we show the predictions for $n_\mathrm{p} = 100$ and $t_\mathrm{p} = 2200, 2300, 2400$ in Fig. 5, where $\hat{y}_{\mathrm{MSE_{min}}}$, $\hat{y}_{\mathrm{MS}}$, $\hat{y}_{\mathrm{MK}}$, $\hat{y}_{\mathrm{ref}}$ are the predictions using N_{MSE}, N_{MS}, N_{MK} and N_{ref}, respectively. From the result for $t_\mathrm{p} = 2300$, we can see that the prediction error by $\hat{y}_{\mathrm{MS}}$, $\hat{y}_{\mathrm{MK}}$ and $\hat{y}_{\mathrm{ref}}$ increases exponentially after $t = 2350$ and the resultant MSE are very big so that it dominates the mean value. From Fig. 4, we can see the mean value of $\mathrm{MSE_{MS}}$ and $\mathrm{MSE_{MK}}$ are smaller than $\mathrm{MSE_{ref}}$ for every horizon n_p. This result indicates the usefulness of the moments of predictive deviations. The result also indicates that we cannot tell whether the present method works or not for a given start time $t_{\rm p}$ and horizon $n_{\rm p}$, but the method works on average.

4 Conclusion

We have analyzed to show that the moments of predictive deviations as ensemble diversity measures can be used for model selection in time series prediction. From the fourth power of bagging ensemble prediction error, we have shown the effect of the moments of predictive deviations of ensemble members to the ensemble prediction error. By means of a number of numerical experiments, we have shown some properties of the moments and utilized the findings for model selection. The method is effective on average but it cannot to be assured for a given start time and horizon of prediction. However, this problem may be related to the exponential prediction error of the chaotic time series prediction, and it is difficult to be solved. We would like to examine this problem and other time series in our future research studies.

This work was supported by JSPS KAKENHI Grant Number 24500276.

References

1. Brown, G., Wyatt, J., Tino, P.: Managing diversity in regression ensembles, J. Mach. Learn. Res. 6, 1621–1650 (2005)
2. Chen, H.: Diversity and Regularization in Neural Network Ensembles. PHD thesis, University of Birmingham (2008)
3. Ono, K., Kurogi, S., Nishida, T.: Moments of predictive deviations for ensemble diversity measures to estimate the performance of time series prediction. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) ICONIP 2012, Part V. LNCS, vol. 7667, pp. 59–66. Springer, Heidelberg (2012)
4. Breiman, L.: Bagging predictors. Machine Learning 26(2), 123–140 (1996)
5. Kurogi, S.: Improving generalization performance via out-of-bag estimate using variable size of bags. J. Japanese Neural Network Society 16(2), 81–92 (2009)
6. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. of the Fourteenth International Conference 18 on Artificial Intelligence (IJCAI), pp. 1137–1143 (1995)
7. Efron, B., Tbshirani, R.: Improvements on cross-validation: the .632+ bootstrap method. J. American Statistical Association 92, 548–560 (1997)
8. Aihara, K.: Theories and applications of chaotic time series analysis. Sangyo Tosho, Tokyo (2000)

A Appendix: Prediction Tools

A.1 Holdout Method for Time Series Prediction

This method is often used to estimate the performance of prediction for unknown time series because it requires only training dataset as follows. For a given time series $y_{t_{\rm g}:n_{\rm g}}$ to predict the successive horizon $y_{t_{\rm g}+n_{\rm g}:n_{\rm p}}$, the former part $y_{t_{\rm g}:n_{\rm g}-n_{\rm p}}$ is used for training and the latter part $y_{t_{\rm g}+n_{\rm g}-n_{\rm p}:n_{\rm p}}$ are hold out to evaluate the prediction performance for model selection.

A.2 CAN2

The CAN2 (competitive associative net 2) is a neural net for learning efficient piecewise linear approximation of nonlinear function by means of the following schemes (See [5] for details): A single CAN2 has N units. The jth unit has a weight vector $\boldsymbol{w}_j \triangleq (w_{j1}, \cdots, w_{jk})^T \in \mathbb{R}^{k\times 1}$ and an associative matrix (or a row vector) $\boldsymbol{M}_j \triangleq (M_{j0}, M_{j1}, \cdots, M_{jk}) \in \mathbb{R}^{1\times(k+1)}$ for $j \in I^N \triangleq \{1, 2, \cdots, N\}$. The CAN2 after learning the training dataset $D^n = \{(\boldsymbol{x}_i, y_i) | y_i = r(\boldsymbol{x}_i) + e_i, i \in I^n\}$ approximates the target function $r(\boldsymbol{x}_i)$ by $\widehat{y}_i = \widetilde{y}_{c(i)} = \boldsymbol{M}_{c(i)}\widetilde{\boldsymbol{x}}_i$, where $\widetilde{\boldsymbol{x}}_i \triangleq (1, \boldsymbol{x}_i^T)^T \in \mathbb{R}^{(k+1)\times 1}$ denotes the (extended) input vector to the CAN2, and $\widetilde{y}_{c(i)} = \boldsymbol{M}_{c(i)}\widetilde{\boldsymbol{x}}_i$ is the output value of the $c(i)$th unit of the CAN2. The index $c(i)$ indicates the unit who has the weight vector $\boldsymbol{w}_{c(i)}$ closest to the input vector $\boldsymbol{x}_i$, or $c(i) \triangleq \underset{j \in I^N}{\text{argmin}} \|\boldsymbol{x}_i - \boldsymbol{w}_j\|$. The above function approximation partitions the input space $V \in \mathbb{R}^k$ into the Voronoi (or Dirichlet) regions $V_j \triangleq \{\boldsymbol{x} \mid j = \underset{i \in I^N}{\text{argmin}} \|\boldsymbol{x} - \boldsymbol{w}_i\|\}$ for $j \in I^N$, and performs piecewise linear prediction for the function $r(\boldsymbol{x})$.

Adaptive Multiple Component Metric Learning for Robust Visual Tracking

Behzad Bozorgtabar[1] and Roland Goecke[1,2]

[1] Vision & Sensing, HCC Lab, ESTeM, University of Canberra
[2] IHCC, RSCS, CECS, Australian National University
Behzad.Bozorgtabar@canberra.edu.au, roland.goecke@ieee.org

Abstract. In this paper, we present a new robust visual tracking approach that incorporates an adaptive metric learning in a multiple components framework. Using a similar overall approach to other state-of-the-art tracking methods, which pose object tracking as a binary classification problem, we firstly employ a new feature selection mechanism based on adaptive metric learning for constructing a discriminative target appearance model and then propose a scheme to update the appearance model in a Multiple Component Learning boosting manner, which automatically learns individual component classifiers and combines these into an overall classifier. Experiments on several challenging benchmark video sequences demonstrate the effectiveness and robustness of our proposed method.

Keywords: Histogram of Oriented Gradients, Adaptive Metric Learning, Multiple Component Learning, Boosting.

1 Introduction

Object tracking is a well-studied problem in computer vision and has many practical applications. The problem and its difficulty depend on several factors, such as the amount of prior knowledge about the target object. Tracking generic objects has remained challenging because an object can drastically change appearance when deforming (e.g. a pedestrian), rotating out of plane, being occluded, or when the illumination of the scene changes.

In *Multiple-Instance Learning (MIL)* [1–3], an object is represented as a bag consisting of a set of feature vectors called instances. In the training set, the labels of bags, either positive or negative, are given, while the uncertainty stems from the unknown labels of instances in the bags. MIL can handle such ambiguity by minimising the negative log likelihood of the training bags, so that a more robust learner can be achieved. However, MIL-based tracking [2] still employs an exhaustive feature selection mechanism to form the adaptive appearance model, which has a negative impact on the accuracy of the tracking system.

In this paper, we present a novel adaptive appearance model and updating method under the recently proposed *Multiple Component Learning (MCL) boosting* framework [4]. Our main contribution is the proposal of an optimisation scheme for the robust updating of the appearance model. Section 2 describes the related work. Section 3 provides an overview of the proposed tracker and its theoretical foundations. The novel

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 566–575, 2013.

MCL-based visual tracking approach is proposed in Section 4. Results on the experimental validation of the approach are presented and discussed in Section 5, before the conclusions are drawn in Section 6.

2 Related Work

The appearance model is an essential part of a tracking system. The question of how to design a robust appearance model, which can be adaptive to the factors mentioned above, is a key task in most recently proposed algorithms [5, 6]. In general, the recently proposed tracking approaches can be categorised into two classes based on their different appearance representation schemes: *generative* models [7] and *discriminative* ones [8, 9]. Generative models first learn an appearance model to represent the object and then search for the object appearance at each frame most similar to the learnt appearance. Black *et al.* [10] learned a subspace appearance model offline. However, the offline learnt appearance model is difficult to adapt to the appearance variations.

To deal with appearance variations, some online learning approaches have been proposed such as the IVT method [7]. Adam *et al.* [11] utilised multiple fragments to design an appearance model, which is robust to partial occlusions. However, these generative models do not take into account background information, eliminating some very useful information that can help to discriminate the object from background [12].

Discriminative models, which are also called tracking-by-detection methods, consider tracking as a binary classification task that separates the object from its surrounding background. Adaptive tracking-by-detection methods first train a classifier in an online manner using samples extracted from the current frame. In the next frame, a sliding window is then used to extract samples around the previous object location, before the previously trained classifier is applied to these samples. The location of the sample with the maximum classifier score is the new object location at the current frame. Collins *et al.* [13] demonstrated that selecting discriminative features in an online manner can greatly improve the tracking performance. Inspired by the advances in face detection [1], many boosting feature selection methods have been proposed. Grabner *et al.* [14] proposed an online boosting feature selection method motivated by the online ensemble method [15]. However, all the above mentioned discriminative methods only utilise one positive sample. If the object location detected by the current classifier is not precise, the extracted positive sample will be imprecise, leading to a suboptimal updated classifier. The inaccuracy will be accumulated to degrade the classifier seriously. Finally, this can lead to tracking failure (due to drift) [2].

Viola *et al.* [1] and Babenko *et al.* [2] introduced the use of MIL for object detection and tracking. Recently, Dollar *et al.* [4] proposed a discriminative learning approach by combining boosting with weakly supervised learning, especially the MIL.

3 System Overview

A typical object tracking system contains three components: an image representation, an object appearance model and a motion model. Given the recent success of the Histogram of Oriented Gradient (HOG) [16] feature in object detection [17], we employ

Algorithm 1. System overview

```
Input: New video frame number $t$
Crop a set of image patches $X^s = \{x \,|s > \|l(x) - l_{t-1}^*\|\}$;
Compute several HOG components feature vectors for each candidate patch in $X^s$;
Use adaptive MCL classifier $\Im(x)$ to estimate $p(y = 1 \,|x)$ for $x \in X^s$;
Update tracked object location $l_t^* = l(\arg\max_{x \in X^s} p(y \,|x))$;
Crop positive samples within a search radius $\alpha$;
Crop two different bags of image patches $X^\gamma = \{x \,|\gamma > \|l(x) - l_t^*\|\}$ and
$X^{\gamma,\beta} = \{x \,|\beta > \|l(x) - l_t^*\| > \gamma\}$, then obtain HOG feature bags for each component
(see Figure 1);
Update object appearance model using the positive patches bag $X^\gamma$ and negative patches
bags $X^{\gamma,\beta}$;
```

a rectangle feature ('patch') as the image representation to describe the components of tracking object. Our appearance model adopts the philosophy of representing the object as an assembly of components [18, 19], but similar to [20], we use a bag of HOG features to model a component. We adopt it in our weak classifier design and evaluate HOG by integral histogram computation [21]. We use the bag of HOG features to model a component and apply a boosting framework to combine the HOG components into a strong discriminant classifier, which is able to return $p(y = 1 \mid x)$ where x is an image patch and y is a binary variable indicating whether x is the target. For the motion model, suppose at time step $t - 1$, our tracker maintains the object location l_t^*:

$$p\left(l_t^* \mid l_{t-1}^*\right) = \begin{cases} 1, \text{ if } \left\|l_t^* - l_{t-1}^*\right\| < s \\ 0, \qquad \text{otherwise} \end{cases} \tag{1}$$

Assuming a trained feature set and labelled samples are given, the proposed method provides an adaptive metric for better tracking. This is especially useful when tracking the target in a rather cluttered environment. Our approach iteratively updates each feature by minimising the weighted least square error between the estimated feature response and the true label.

The basic flow of our tracking system is illustrated in Figure 1. The main procedures of our system are as follows: Let l_t^* denote the location of the object at the t^{th} frame. First, we densely crop some positive samples $X^\alpha = \{x \mid \alpha > \|l(x) - l_*^t\|\}$ within a search radius α centring at object location l_t^* and then we crop the positive bag of image patches from set $X^\gamma = \{x \mid \alpha > \|l(x) - l_*^t\|\}$. Second, we randomly crop some negative samples from set $X^{\gamma,\beta} = \{x \,|\beta > \|l(x) - l_t^*\| > \gamma\}$ where $\alpha < \gamma < \beta$. Third, we utilise these positive and negative bags to update the classifier $\Im(x)$. In the next frame, we crop some samples $X^s = \{x \mid s > \|l(x) - l_t^*\|\}$ with a large radius surrounding the previous object location at the $t + 1^{th}$ frame. Next, we apply the previously trained classifier to these samples to find the sample with the maximum confidence. Based on the newly detected object location, our tracking system repeats the above mentioned procedures. The overview of our tracking system is summarised in Algorithm 1.

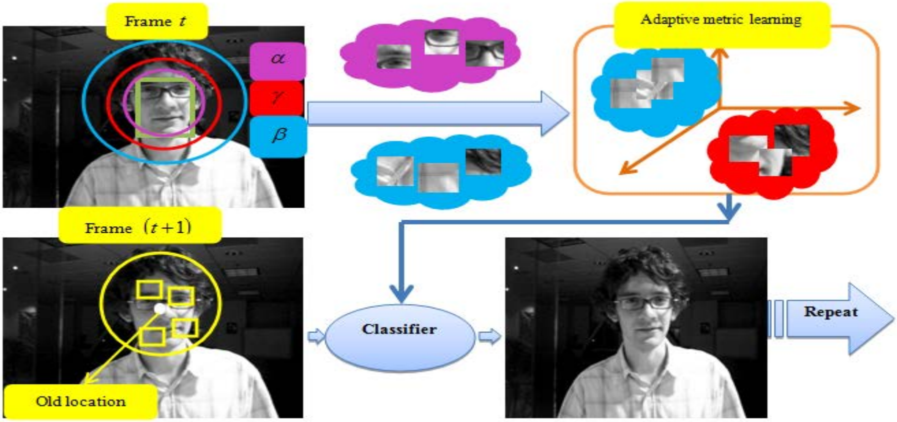


Fig. 1. The basic flow of our system

3.1 Adaptive Metric Learning

Choosing an appropriate distance measure is fundamental to the tracking problem, especially when the target is subject to significant variation such as changes in illumination. To measure the neighbourhood relation between a pair of feature vectors x_i and x_j in the training set, we use p_{ij} as a soft-max over the Euclidean distance similar to [22] in the new feature space transformed by a linear transformation matrix $A \in \Re^{d\times m}$ where d is the dimension of the transformed feature space, set to 12 in the experiments, and m is the dimension of the input feature space, set to 144:

$$p_{ij} = \frac{\exp\left(-\left\|Ax_i - Ax_j\right\|^2\right)}{\sum_{k\neq i}\exp\left(-\left\|Ax_i - Ax_k\right\|^2\right)} \quad . \tag{2}$$

Assuming all instances in the negative bags are correctly labelled and that instance labels in the positive bag are unknown, we aim to maximise the objective function $g\left(A\right)$, which is the expected number of negative samples in the input feature space that are correctly classified [22]:

$$A^* = \arg\max_{A} g\left(A\right) \tag{3}$$

$$= \arg\max \sum_{i=-1} \log\left(\sum_{j\in C_{i=-1}} p_{ij}\right) \quad . \tag{4}$$

The above optimisation problem can be solved by a gradient descend technique [22], in order to compute $\frac{\partial g}{\partial A}$:

$$\frac{\partial g}{\partial A} = 2A\sum_{i=-1}\left(\sum_{k} p_{ik}x_{ik}x_{ik}^T - \frac{\sum_{j\in C_{i=-1}} p_{ij}x_{ij}x_{ij}^T}{\sum_{j\in C_{i=-1}} p_{ij}}\right) \tag{5}$$

where $x_{ij} = x_i - x_j$.

3.2 Component Classifier

The Histogram of Gradient feature is robust to pose variation [16] and is widely applied in tracking applications. We use a 9-bin histogram of gradient magnitude at each orientation for the blocks containing 4×4 cells. Note that the HOG feature can be parametrised by (x_0, y_0, x_1, y_1), where (x_0, y_0) and (x_1, y_1) are the two corners (top-left and bottom-right) of the first cell. We adopt it in our weak classifier design and evaluate HOG components by integral histogram computation [21]. After transforming features to the new feature space, the component classifier according to the HOG feature can be achieved by using SVM classifier. Therefore, the component classifier can be presented as an SVM of:

$$F_m \left(x; \mathbf{p}\right) = A \left(x\right)^T \alpha + b \tag{6}$$

where x denotes the candidate image patch. The parameters of component classifier $\mathbf{p} = [A, \alpha, b]^T$ and M denotes the number of components for representing the target $(m = 1, \cdots, M)$.

3.3 Feature Selection

We aim to adaptively select features that are of the most discriminative ability from the pool. Then, the feature selection can be seen as a process of updating the parameters of each weak classifier. Therefore, it is natural to use the weighted least square error (WLSE) as the objective function for feature updating:

$$\min \varepsilon \left(F \left(x;\ \mathbf{p}\right)\right) = min \sum_{i=1}^{K} D\left(i\right) \left(F\left(x; \mathbf{p}\right) - y_i\right)^2\Big) \tag{7}$$

where y_i denotes the label of the i^{th} training image patch. Inspired by the definition of Multiple Instance Learning, in our proposed framework, we use training data bags $\{(X_1, y_1), \ldots\}$ where $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}\}$, x_{ij} denotes an image patch and y_i a pixel location. Similar to [2], we use a boosting framework to train a boosted classifier to minimise the negative log likelihood of bags:

$$L = -\sum_{i} \left(\log p\left(y_i \left| X_i \right.\right)\right) \tag{8}$$

where $p\left(y_i \left| X_i \right.\right)$ is the posterior probability of the bag, which we denote by p_i.

Moreover, we should define the posterior probability $p\left(y_{ij} \mid x_{ij}\right)$ of an instance at bag X_i and take the p_{ij}. Similar to the definition of the bag label, the connection between bag probability p_i and the probability of its instance p_{ij} can be achieved by:

$$p_i = max_j \left(p_{ij}\right) \tag{9}$$

Several differentiable approximations to the max operator exist in the literature. In [1], the Noisy-OR (NOR) model is adopted for doing this: Our proposed framework iteratively minimises the negative log likelihood of training bags by choosing the most discriminative feature from a large feature pool in each boosting training phase.

$$p_i = 1 - \prod_j (1 - p_{ij}) \tag{10}$$

In order to measure the contrast between the confidence that one sample would be classified as positive or negative, we use $h_m(x)$ (the log odd ratio of weak classifier) to model the instance probability instead of $F_m(x; \mathbf{p})$ itself directly. The $h_m(x)$ can be represented as:

$$h_m(x) = log\left[\frac{p(y = 1 \mid F_m(x))}{p(y = -1 \mid F_m(x))}\right] \quad . \tag{11}$$

For each trained SVM $F_m(x; \mathbf{p})$, we compute the mean μ^+ and μ^- of positive and negative support vectors [23], respectively. Then, $p(y = 1 \mid F_m(x))$ and $p(y = -1 \mid F_m(x))$ are computed as follows:

$$p(y^* \mid F_m(x)) = exp(-|A(x) - \mu_m^*|) \tag{12}$$

where $y^* = 1$ or $y^* = -1$ when μ^* is μ^+ or μ^-. Then, the instance probability can be modelled as:

$$p(y_i \mid x_{ij}) = \sigma(H_{m-1}(x_{ij}) + h_m(x_{ij})) \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid function, $H_{m-1}(x)$ is the sum of the log odd ratio of the previous component classifiers. Finally, the bag probability $p(y_i \mid X_i)$ is modelled using a Noisy-OR model.

4 Multiple Component Metric Learning

We introduce a novel visual tracking algorithm that builds a discriminative appearance model of the target object via adaptive metric learning for the feature selection and updates such a model based on multiple components, then optimises them under an online boosting framework. It inherits the essential idea from MCL [4] and initial definitions from MIL tracking [2]. Algorithm 2 presents the pseudo-code of the online *Multiple Component Metric Learning (MCML)*. The likelihood function L^m is computed at each iteration and expected to decrease monotonically.

5 Experiments and Discussion

Experimental Settings: We evaluated the proposed *MCML* tracking algorithm on five challenging benchmark video sequences, all of which are publicly available [2]. In addition, we tested three other visual trackers including *Online-AdaBoost (OAB)* [14], *Fragment Tracking (FragTrack)* [11] and *MIL Tracker (MIL)* [2] on the same video sequences for comparison. All algorithm parameters are fixed for all test video sequences.[1] In addition, the number of candidate weak classifiers in the feature pool is set to 250 for OAB and MIL ($\alpha = 3$).

[1] The number of component classifiers M is set to 50. For our method, positive samples are cropped from all positions within a radius of $\gamma = 5$ pixels, while negative samples are cropped between γ and twice the size of the target patch by random sampling.

Algorithm 2. Adaptive Multiple Component Metric Learning

```
   Input  : Dataset $\{X_i, y_i\}_{i=1}^{N}$, where $X_i = \{x_{i1}, x_{i2}, \ldots\}$, $y_i \in \{1, -1\}$ and initial set of
            component classifiers
1  Initialise $\Im(x) = 0$, weights $D_1(i) = \frac{1}{K}$;
2  for $m = 1$ to $M$ do
3      Extract HOG features for current component;
4      Calculate $g(A)$, $\frac{\partial g}{\partial A}$ and $\varepsilon_m = \sum_{i=1}^{K} D_m(i) \|F_m(x; c_m) - y_i\|$ where $y_i$ is the
       class label of the $i^{th}$ training image patch;
5      $L^m = -\left(\sum_{i|y_i=1} \log(p_i^m) + \sum_{i|y_i=-1} \log(p_i^m)\right)$;
6      if $L^m$ is decreasing then
7          Update transformation matrix for new components;
8          Go to step 3;
9      end
10     Update $\Im(x) = \Im(x) + \alpha_m . F_m(x; c_m)$;
11     Update the weights
12     $D_m(i) = D_m(i) . \exp(-\alpha_m y_i F_i) / Z_m$, where $\alpha_m = \frac{-1}{2} \log\left(\frac{\varepsilon_m}{1-\varepsilon_m}\right)$ and
       $Z_m = 2\sqrt{\varepsilon_m . (1 - \varepsilon_m)}$ is the normalisation factor.
13 end
   Output: The strong classifier:
           $\Im(x) = sign\left(\sum_{m=1}^{M} F_m(x; c_m)\right)$ (see Figure 2)
```

Quantitative and Qualitative Analysis: The ground truth of the centre position of target objects in the video sequences labelled every five frames are provided by Babenko *et al.* [2]. All testing video frames are gray scale and resized to 320 × 240 pixels. We use the average centre location error as the evaluation criterion to compare performance. In Table 1, each row represents the average centre location errors of the four algorithms tested on a certain video sequence. **NaN** denotes the tracker lost the target for several frames. Figure 3 shows some example tracking results on example video frames from the Tiger and Face Occlusion videos, respectively.

1. Videos: Occluded Face & Girl

The *Occluded Face* video is designed to evaluate whether a tracking algorithm can handle partial occlusion as well as pose changes, e.g. object rotation in the plane. Only

Table 1. Average centre location errors (in pixels). NaN refers to tracking being lost.

Video Clip	Fragment	OAB	MIL	**MCML**
Occluded Face	30	31	15	**11**
Girl	32/**NaN**	58	30	**13**
David Indoor	70	28	22	**12**
Coke Can	38	10/**NaN**	15	**14**
Tiger 1	49/**NaN**	45	23	**10**

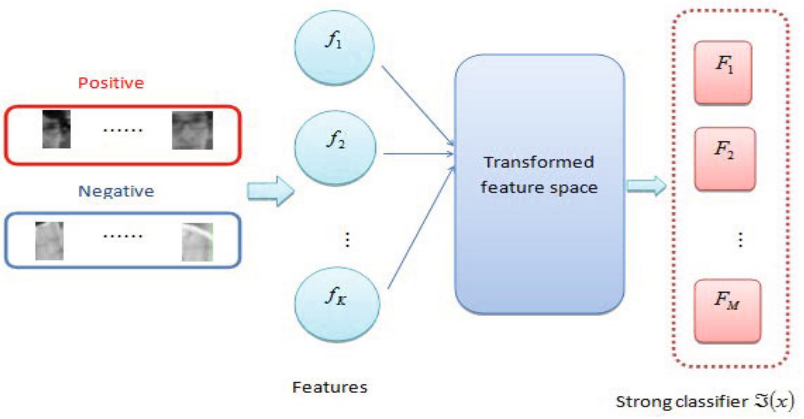


Fig. 2. Illustration of boosting process

the results of our approach and MIL tracker are broadly comparable. However, the MIL tracker fails to handle partial occlusion as can be seen in Figure 3 (right). The *Girl* sequence is a good example of out-of-plane rotation. The error of our proposed MCML tracker is less than half of that of other methods.

2. Video: David Indoor

This video sequence is widely used as a benchmark in state-of-the-art tracking systems, e.g. [2]. It includes challenging changes in illumination, scale and pose. Our proposed MCML tracker clearly outperforms the second best tracker (MIL) in this video sequence in terms of average centre location errors in pixels as can be seen in Table 1. One of the advantages of our method is in using HOG features, which are invariant to pose changes.

3. Videos: Tiger 1 & Coke Can

Of all the testing videos, the Tiger sequence includes the most challenges such as frequent occlusions, fast motion, motion blur, and drastic changes in object appearance. Our proposed MCML tracker again clearly outperforms all other methods with an average centre location error of only 10 pixels. *FragTrack* loses the target several times after frame #155, while MIL and OAB lose the target in some frames. OAB has the lowest error rate overall in the Coke Can sequence, but loses track at one point. Our proposed MCML tracker has a slightly higher error but continuously tracks the object.

6 Conclusions

In this paper, we have proposed a novel visual tracking framework based on adaptive metric learning, which inherits its idea from Multiple Component Learning. The proposed algorithm not only achieves a suitable way of updating the discriminative feature set using the adaptive metric scheme, but overcomes the drifting problem with the help of MIL. The experimental validation on several difficult benchmark videos demonstrates the performance of our proposed method.

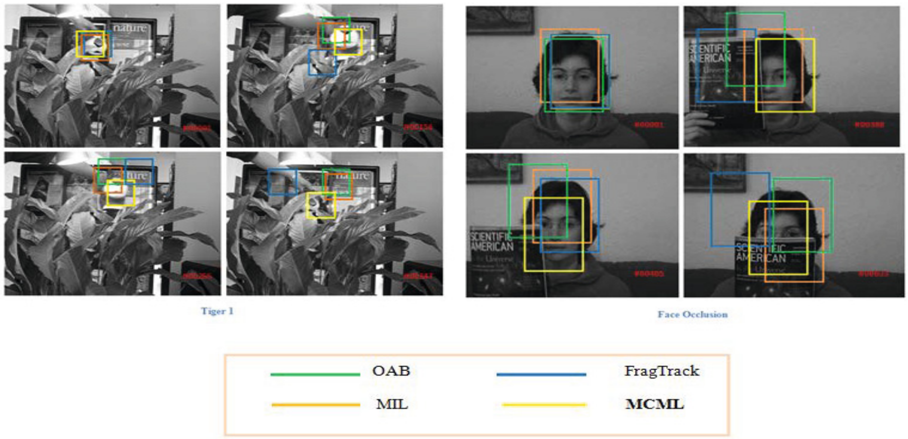


Fig. 3. Example tracking results for selected frames from the Tiger (left) and Face Occlusion (right) sequences

References

1. Viola, P., Platt, J., Zhang, C.: Multiple Instances Boosting for Object Detection. In: Neural Information Processing Systems, pp. 1417–1426 (2005)
2. Babenko, B., Yang, M., Belongie, S.: Visual Tracking with Online Multiple Instance Learning. In: Computer Vision and Pattern Recognition, CVPR (2009)
3. Dietterich, T., Lathrop, R., Perez, L.: Solving the Multiple Instance Problem with Axis-Parallel Rectangle. Artificial Intelligence 89, 31–71 (1997)
4. Dollár, P., Babenko, B., Belongie, S., Perona, P., Tu, Z.: Multiple Component Learning for Object Detection. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 211–224. Springer, Heidelberg (2008)
5. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(10), 1296–1311 (2003)
6. Mei, X., Ling, H.: Robust visual tracking using? 1 minimization. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1436–1443. IEEE (2009)
7. Ross, D., Lim, J., Lin, R.S., Yang, M.H.: Incremental Learning for Robust Visual Tracking. International Journal of Computer Vision 77(1), 125–141 (2008)
8. Avidan, S.: Ensemble tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(2), 261–271 (2007)
9. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: Proc. BMVC, vol. 1, pp. 47–56 (2006)
10. Black, M.J., Jepson, A.D.: Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. International Journal of Computer Vision 26(1), 63–84 (1998)
11. Adam, A., Rivlin, E.: Robust Fragments-based Tracking Using the Integral Histogram. In: Computer Vision and Pattern Recognition (CVPR), pp. 798–805 (2006)

12. Wang, Q., Chen, F., Xu, W., Yang, M.H.: An experimental comparison of online object-tracking algorithms. In: SPIE Optical Engineering+ Applications, International Society for Optics and Photonics, pp. 81381A–81381A (2011)
13. Collins, R.T., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1631–1643 (2005)
14. Grabner, H., Grabner, M., Bischof, H.: Real-Time Tracking via Online Boosting. In: British Machine Vision Conference (BMVC), pp. 47–56 (2006)
15. Oza, N.C., Russell, S.: Online ensemble learning. University of California, Berkeley (2001)
16. Dalal, N., Triggs, B.: Histogram of Oriented Gradient for Human Detection. In: Computer Vision and Pattern Recognition (CVPR), pp. 886–893 (2005)
17. Laptev, I.: Improvements of object detection using boosted histograms. In: BMVC (2006)
18. Fergus, R., Perona, P., Zisserman, A.: A sparse object category model for efficient learning and exhaustive recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 380–387. IEEE (2005)
19. Leibe, B., Schindler, K., Cornelis, N., Van Gool, L.: Coupled object detection and tracking from static cameras and moving vehicles. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10), 1683–1698 (2008)
20. Xie, Y., Qu, Y., Li, C., Zhang, W.: Online multiple instance gradient feature selection for robust visual tracking. Pattern Recognition Letters (2012)
21. Porikli, F.: Integral histogram: A fast way to extract histograms in cartesian spaces. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 829–836. IEEE (2005)
22. Jiang, N., Liu, W., Wu, Y.: Learning Adaptive Metric for Robust Visual Tracking. IEEE Transactions on Image Processing 20, 2288–2300 (2011)
23. Yang, F., Lu, H., Chen, Y.-W.: Human tracking by multiple kernel boosting with locality affinity constraints. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part IV. LNCS, vol. 6495, pp. 39–50. Springer, Heidelberg (2011)

Empirical Evaluation on Deep Learning of Depth Feature for Human Activity Recognition

Junik Jang, Youngbin Park, and Il Hong Suh

Electronics and Computer Engineering, Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, Korea
junik2020@gmail.com {pa9301,ihsuh}@hanyang.ac.kr

Abstract. In the field of computer vision, there are two emerging approaches that have drawn much attention, and they have recently become popular way to solve various kinds of recognition problem. The first approach is unsupervised feature learning based on deep learning technique, and second approach is to conduct recognition using depth information thank to recent progress in depth sensor. At this point, it seems reasonable that one is curious about effectiveness of deep learning from raw depth data. However, a few researches have attempted to learn depth features with a deep network, and the validity has not been well studied in terms of quantitative analysis. To this end, we learned depth features for human activity recognition using existing deep learning algorithm and evaluated effectiveness of the learned depth feature on activity recognition. Furthermore, we provide analysis in detail and valuable discussion with additional experiments.

Keywords: deep learning, human activity recognition, Kinect sensor.

1 Introduction

In the field of computer vision, two emerging approaches have drawn much attention, and they have recently become popular way to solve various kinds of recognition problem.

First approach is unsupervised feature learning based on deep learning technique such as Deep Belief Nets [1], deep Boltzmann machines [2], convolutional deep belief networks [3] and Stacked Autoencoders [4]. In particular, they construct deep feature representation by learning features layer by layer, in which the features are learned directly from raw vision data in unsupervised manner. For this reason, deep learning is generalizable while hand-designed features such as SIFT [5] and HOG [6] are not easily extended to other sensor modalities. Furthermore, several studies have revealed that deep learning not only generalizes to different domains but also achieves impressive performance on many types of recognition tasks such as handwritten digit recognition [7], scene recognition [8], object recognition [9], and human activity recognition [10]. Therefore, there is a growing interest in learning feature based on deep learning.

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 576–583, 2013.

Second approach is to conduct recognition using depth information thank to recent progress in RGB-D sensor. Stereo camera and laser sensor had been commonly used to acquire depth cue before Microsoft introduced a RGB-D camera called Kinect. Microsoft Kinect is novel sensing systems that capture RGB images along with per-pixel depth information and is regarded as more efficient device than stereo camera and laser sensor because of its frame rate, low cost, and accuracy. Analogous to deep learning, many researchers recently employ Kinect for recognition and demonstrate good performance [11], [12], [13].

At this point, it seems reasonable that one is curious about effectiveness of deep feature learning from raw depth data. However, a few researches [14] have attempted to train depth features based on a deep unsupervised learning algorithm, and the validity has not been well studied in terms of quantitative analysis. To this end, we learned depth features for human activity recognition using existing deep learning algorithm, and evaluated the performance of the learned features on the activity recognition. In fact, activity recognition is usually performed well by skeleton data provided by Kinect SDK, but in this paper, we do not consider such hand-designed features.

Specifically, we employed a model developed by Le. et al. [10] to learn depth feature. The model achieves state-of-the-art performance on several human activity dataset, it builds feature hierarchy based on deep learning techniques, in which spatio-temporal features are learned directly from RGB video. To learn depth feature, first, we captured several videos by a Kinect. Captured RGB-D videos was then divided into depth and RGB channels, and the two channels were separately given as inputs for two identical deep learning processes. Consequently, we had two feature hierarchies for depth and RGB cues, respectively. We followed standard processing pipeline described in [15] for subsequent process, in which after extracting local features based on a learned feature hierarchy, vector quantization is performed by K-means, and conduct classification with non-linear SVMs.

2 Spatio-temporal Feature Learning

As mentioned earlier, we employed a model developed by Le. et al. [10] to build feature hierarchy for depth cue. The model extends Independent Subspace Analysis (ISA) [16] algorithm by combining ISA with deep learning techniques such as stacking and convolution. In this section, we will briefly describe Independent Subspace Analysis algorithm, how to combine ISA with deep learning techniques, and spatio-temporal feature learning based on depth information. More details are presented in [10].

2.1 Independent Subspace Analysis

In fact, ISA is extension of Independent Component Analysis (ICA) [17]. An advantage of ISA, compared to the ICA algorithm is that it learns robust feature to local translation. As shown in Figure 1. an ISA is a two-layered network, in

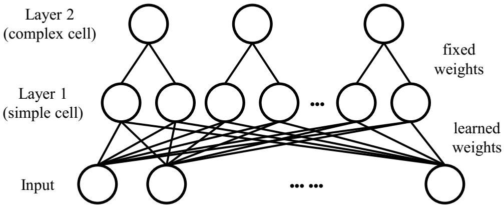


Fig. 1. The architecture of an ISA. It is a two-layered network.

which first and second layers consists of simple and complex cells, respectively. The weight W between input and simple cells are learned, and the weights V between simple cells and complex cells are fixed. Each of the complex cells pools over a small neighborhood of adjacent simple cells.

The activation of a complex cell is given as

$$p_i(x^t; W, V) = \sqrt{\sum_{k=1}^{m} V_{ik} (\sum_{j=1}^{n} W_{kj} x_j^t)^2} \tag{1}$$

where x_j^t is an input pattern. According to this equation, Responses of both simple and complex cells are determined by a nonlinear function that follow a linear stage. Square and square-root functions are employed for simple and complex cells, respectively. The objective function to learn model parameters W is defined as

$$\begin{aligned} \underset{W}{\text{minimize}} \quad & \sum_{t=1}^{T} \sum_{i=1}^{m} p_i(x^t; W, V) \\ \text{subject to} \quad & WW^T = I. \end{aligned} \tag{2}$$

where $\{x^t\}_{t=1}^{T}$ indicates all training data, and m is the number of complex cells. The constraint is to ensure the features are diverse. After training ISA, Gabor filter like features with many frequencies and orientations are learned.

2.2 Hierarchical ISA

A disadvantage of ISA, analogous to ICA is that training can be very slow when the dimension of the input data is large. This is because an orthogonalization step in Equations 2. The cost of the step grows as a cubic function of the input dimension. To learn spatio-temporal features for human activity recognition, ISA thus should be scaled up to deal with large dimension of input. For this purpose, a ISA is first trained on small input patches and first layer is composed of the ISA. After training layer 1, this learned network then is taken and is convolved with a larger region of the input image. The responses of this convolution step are given as input to train next layer ISA. This convolution and stacking step can be repeated to learn a hierarchical feature representation.

2.3 Learning Spatio-Temporal Feature from Depth Cue

In contrast to object recognition from static scene, when we try to recognize human activity from video, spatio-temporal feature is more suitable than 2D feature. It is straightforward to apply the Hierarchical ISA to the video domain. First of all, R, G, and B images at each time of video are converted into a grayscale image. Then, a sequence of image patches instead of a image patch is flattened to a vector, and the vector is given as input to the network in Figure 1.

To learn spatio-temporal feature from depth cue is very similar to this process. We first capture several videos by a Kinect and extract depth channels from the captured RGB-D videos. Then, a sequence image patches from the the depth channel is given as input to the Hierarchical ISA.

3 Experiments

3.1 Datasets and Experimental Setup

We build a two-layered hierarchical ISA. The dimension of input for ISA at layer 1 is 2560, the number of simple and complex cells for ISA at layer 1 are 300 and 150, respectively. The dimension of input for ISA at layer 2 is 1200, the number of simple and complex cells are 300 and 100, respectively.

We captured 245 videos by a Kinect, in which there are 7 activity categories such as turning, waking, extending a hand to a cup, picking up a cup, drinking, putting down a cup, and standing. For a activity class, 19 and 16 videos were used for training and test, respectively. The length of a video is about 60 frames and image resolution is 320 x 240. Captured RGB-D video was then divided into depth and RGB channels, and the two channels were separately given as inputs for two identical deep learning processes.

During training, 2000 video blocks, each block is 20 x 20 spatial size and 14 temporal size, were randomly sampled from each captured video. A vectors given as input for ISA at layer 1 and 2 are selected in the following way: 8 small video blocks, each block is 16 x 16 spatial size and 10 temporal size, are extracted from a video block with partial overlapping. Each small block is flattened to a vector given as input for ISA at layer 1. As a result, total number of responses obtained from a original video is 1200, and they are given as input for ISA at layer 2. This is convolution step as mentioned in Section 2.2.

3.2 Processing Pipeline

We used an standard pipeline as described in [15]. This pipeline extracts local features, then performs vector quantization by K-means and classifies by χ^2-kernel SVM.

More specifically, when a video block sampled from a video is given to the learned ISA hierarchy, responses at layer 1 and 2 are combined, quantized into visual words, and then used for local features. After all video blocks sampled

Table 1. Recognition accuracy

Method	RGB	Depth
Accuracy	95.93	94.38

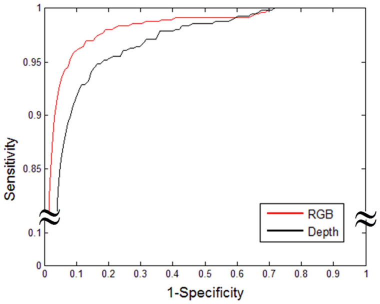


Fig. 2. ROC curves illustrating the performances of the depth and RGB features based recognition

from a video are given to the network, a video is represented as the frequency histogram over the visual words.

The vocabularies are constructed with k-means clustering. The number of visual words is 4000 which has shown to empirically give good performance for a wide range of datasets. We implemented the pipeline with an extended SVM proposed by [18], which can give probability estimates. For multi-class classification, one-versus-rest approach was employed.

3.3 Comparison of Depth and RGB Channels Based Recognitions

We employ two standard measures for numerical evaluation, which are accuracy and ROC curve. Accuracy is (# of true positives + # of true negatives)/(# of test data). Because we use one-versus-rest approach for classification, each SVM for a action category provides probability in terms of positive and negative. For calculation of accuracy, we set threshold to 0.5. The accuracy is averaged over 7 activity categories.

Table 1 shows the results, in which accuracy of RGB channel based method is about 1.5 higher than that of depth. ROC curve is shown in Figure 2. Two quantitative measures demonstrate that RGB channel based recognition achieves superior performance compared to depth channel based approach.

4 Discussions

The evaluation results demonstrate that the learned depth features based on deep learning is not effective compared to the learned RGB features. However, still, some questions are remained as follows:

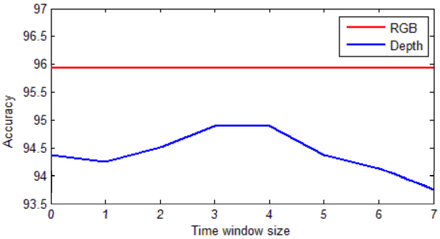


Fig. 3. Accuracy curve by varying the size of time window

- Why is the depth features inferior to RGB features?
- Even though depth features is inferior to RGB features, are the depth features good complements to the RGB features?

We assume the reason of first question is low quality of depth cue. In fact, in spite of many advantages of Kinect, quality of depth cue obtained from Kinect is not good as that of RGB cue. Kinect often fails to obtain measurements in some areas, and these appear as small and large holes in a depth image. This is mainly caused by lighting, occlusion, objects being out of range, and objects absorbing rather than reflecting infrared. Even worse, positions of Kinect and objects are fixed, edges of depth image flicker. We consider flickering as more important issue. The flickering might lead bad effects on learned depth features because the unsupervised feature learning from raw depth cue generally produces various kinds of edge detectors.

To verify above assumption, we improved quality of raw depth cue by traditional techniques such as spatial and temporal smoothings. Details are described in [19]. Briefly, spatial smoothing selects depth with highest frequency over neighboring pixels and temporal smoothing determines depth by weighted average of depths in time window. The size of time window in temporal smoothing controls the amount of flickering for fixed objects and afterimage for moving objects. If afterimage is clearly observed this implies that the depths are too much averaged. Therefore, in our case, flickering has trade-off relation with afterimage. As the size is larger, flickering over fixed objects is decreased, but afterimage for moving object is more clearly notable.

Even though remarkable improvement cannot be achieved by the two methods, but we can expect slightly better quality. Since we have a more focus on flickering problem, we varied the size of time window in temporal smoothing, then recorded the recognition performance. Figure 3 shows the results. In is noted that as the size of time window become larger, performance grows gradually in a certain region. This means that the performance has correlation with the quality of raw depth cue such as flickering and small hole. However, it still cannot be decided that the depth features based recognition outperforms the RGB features based results when it allows that we can learn depth features from completely noiseless depth cue, As the size become lager than 4, performance is decreased. We estimate this result due to too much average.

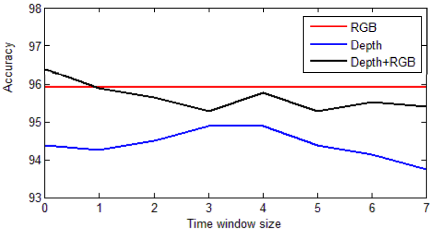


Fig. 4. Accuracy curve by varying the size of time window

To answer second question, the responses from RGB and depth hierarchies have been combined. The combined responses were then quantized into visual words and then used for local features. If the depth features are suitable complements to the RGB features, the accuracy based on the combined features should outperform the accuracy based on the RGB features, otherwise the depth features do not lead complement effects. Black line in Figure 4 shows the performance of the combined features. In a certain region, the performance is little bit higher than that of RGB features, but since the difference is tiny it cannot be regarded that the combined features outperforms the RGB features. One of the possible explanations is redundancy. If the learned depth features contain information mostly extracted from the silhouette of objects, these information might be redundant with the learned RGB features.

5 Conclusions

In this paper, we learned depth features from raw depth frames using existing deep learning algorithm and evaluated effectiveness of the learned depth feature based on activity recognition performance. In fact, activity recognition is usually performed well by skeleton data provided by Kinect SDK, but in this paper, we do not consider such hand-designed features. In experiments, depth channel based recognition demonstrated poor performance compared to RGB channel based approach. Furthermore, we found that the performance has correlation with the quality of raw depth cue such as flickering and small hole and the learned depth features are not suitable complements to the RGB features.

Acknowledgments. This work was supported by the Global Frontier R&D Program on <Human-centered Interaction for Coexistence> funded by the National Research Foundation of Korea grant funded by the Korean Government (MEST) (NRFM1AXA003-2011-0028353). All correspondences should be addressed to I. H. Suh.

References

1. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithms for deep belief nets. Neu. Comp. (2006)
2. Salakhutdinov, R., Hinton, G.: Deep Boltzmann Machines. In: International Conference on AI and Statistics (2009)
3. Lee, H., Grosse, R., Ranganath, R., Ng, A.: Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In: ICML (2009)
4. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layerwise training of deep networks. In: NIPS (2006)
5. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
7. Hinton, G., Osindero, S., Teh, Y.: A Fast Learning Algorithm for Deep Belief Nets. Neural Computation 18(7), 1527–1554 (2006)
8. Bo, L., Ren, X., Fox, D.: Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms. In: NIPS (2011)
9. Yu, K., Lin, Y., Lafferty, J.: Learning Image Representations from the Pixel Level via Hierarchical Sparse Coding. In: CVPR (2011)
10. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis. In: CVPR (2011)
11. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE CVPR (2011)
12. Koppula, H.S., Saxena, A.: Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation. In: ICML (2013)
13. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: ICRA (2012)
14. Socher, R., Huval, B., Bath, B.P., Manning, C.D., Ng, A.Y.: Convolutional-Recursive Deep Learning for 3D Object Classification. In: NIPS (2012)
15. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2010)
16. Hyvarinen, A., Hurri, J., Hoyer, P.: Natural Image Statistics. Springer (2009)
17. Pierre, C.: Independent Component Analysis: a new concept? Signal Processing 36(3), 287–314 (1994)
18. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2003)
19. http://www.codeproject.com/Articles/317974/KinectDepthSmoothing

Salient Object Segmentation Based on Automatic Labeling

Lei Zhou[1], Chen Gong[1], YiJun Li[1], Yu Qiao[1], Jie Yang[1], and Nikola Kasabov[2]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, China
[2] Auckland University of Technology, New Zealand

Abstract. This paper proposes an automatic salient object extraction framework. Firstly, the saliency model are developed by applying the low level color features and the boundary prior. The initial salient regions are extracted by adaptive thresholding. Multiple classifiers are trained with extracted initial region, which reflect color information of images or adopt label propagation. Then, the labels for segmentation are generated automatically via classifier composition. Finally, the conditional random field (CRF) model based on multi-feature fusion is applied for salient object segmentation. Empirical study reveals that the proposed algorithm achieves satisfying performance.

Keywords: saliency detection, automatic object segmentation, automatic labeling, conditional random field.

1 Introduction

Object segmentation is a challenging problem in computer vision and it has wide applications in areas such as object recognition, image classification and image retrieval, etc. Therefore, many methods have been proposed to extract interesting objects automatically. Salient object extraction can be formulated as a binary labeling problem which assigns a unique label to each pixel (belonging to salient object or background), and the labeling problem is often formulated as a minimization of the energy [1]. In the past few years, many energy formulations have been developed which adopt either markov random field (MRF) or conditional random field (CRF). The efficiency of methods mainly lies in how the appearance cues, such as color, texture or valuable high level information, are defined and incorporated into the segmentation model. In the context of salient object segmentation based on saliency map, the key issue is how to utilize the saliency model efficiently. Many works focus on incorporating the saliency information into segmentation model directly. In [2], the saliency map obtained via maximal symmetric surrounding region is directly exploited to construct the data term for graph cut. In [3], saliency map and color similarity are used to define the two complementary data terms and the weights for the two data terms are set adaptively. There also exist works which extract the initial regions of salient objects based on the saliency models, so as to define more discriminative features for

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 584–591, 2013.

segmentation. In [4], the seeds of salient object/background are selected manually by thresholding on the saliency map and they are updated iteratively. In [5], an iterative unsupervised salient object segmentation approach based on kernel density estimation (KDE) and two-phase graph cut is proposed. In [6], a CRF model is constructed to integrate cues such as color and context information. There are also many strategies which extract initial salient regions based on the proposed saliency map of high quality or design schemes that are more robust than thresholding based on existing saliency models. In [7], convex hull analysis is performed on several binary object masks which are generated by diverse saliency maps, to select the most compact shape to represent the object. In [5], an initial segmentation is generated by thresholding on kernel density estimation based saliency model. In [6], an adaptive selection mechanism is designed to select the minimal connected region of the saliency map as the initial segmentation of the salient object according to three measures, connectivity, convexity and saliency.

In order to enhance the segmentation reliability, especially for complicated images, and improve the overall segmentation quality, we propose an efficient saliency model which exploits the low level features, such as color contrast and color distribution, and the boundary prior. Then, a initial segmentation of the salient object is selected according to the saliency map. To extract the initial region more precisely, color Gaussian mixture model (GMM) and a semi-supervised label propagation method are applied to generate seeds automatically via classifier composition. Then, the seeds for salient objects and background are generated automatically, and they are used to train the appearance cues for segmentation. Finally, a CRF model is constructed for obtaining the final segmentation results. The main contributions of the paper are summarized below:

1 An automatic segmentation algorithm that can extract object from background without any interaction is proposed.

2 A saliency model that integrates low level features of images and prior information related to boundaries of images is developed.

3 An automatic labeling scheme based on classifiers composition is presented.

2 Saliency Model Combination

Color Contrast and Color Distribution: Color contrast is inspired by the observation that color components of a salient object may have a strong contrast to their surroundings. Assume that an image is divided into regions (or superpixels) $R_i, i \in \{1, 2, 3, ..., N\}$. Then, region i's color contrast saliency S_i^{con} is computed according to the definition in [8]:

$$S_i^{con} = \sum_{j \neq i} D_c(R_i, R_j) D_s(R_i, R_j), \qquad (1)$$

where $D_c(R_i, R_j)$ is the color distance between the two regions, and $D_s(R_i, R_j)$ is the spatial distance between the regions R_i and R_j.

The distribution of color information in R_i, D_i^{dist} is defined in eq. (2) according to the definition in [8]:

$$D_i^{dist} = \sum_{j \neq i} w_{ij}^C (p_j - \overline{p_i})^2. \tag{2}$$

In eq. (2), p_i describes the average position of superpixel i and $\overline{p_i}$ is the weighted average position. w_{ij}^C is the weight corresponding to color similarity between the region i and region j. The regions with higher distribution variances may have lower saliency, so we define the color distribution saliency as:

$$S_i^{dist} = 1 - D_i^{dist}. \tag{3}$$

Boundary Prior: In an image, the object near to the boundary is less-likely to be the salient object. Geodesic distance is computed based on the nearest background nodes Ω_B which are selected by an method similar to [9]. For the pixel m, the distance is defined as $g(m) = \min_{s \in \Omega_B} d_g(s, m)$. The geodesic distance is computed based on the length of a discrete path:

$$L(\Gamma) = \sum_{i=1}^{n-1} \sqrt{(1-\gamma_g) d(\Gamma^i, \Gamma^{i+1})^2 + \gamma_g \parallel \nabla(\Delta^i) \parallel^2}. \tag{4}$$

where Γ is an arbitrary discrete path with pixels defined as $\{\Gamma^1, ..., \Gamma^n\}$. $d(\Gamma^i, \Gamma^{i+1})$ is the Euclidean distance between two points (Γ^i and Γ^{i+1}). Then the distance is defined as $d_g(a, b) = \min_{\Gamma \in P_{a,b}} L(\Gamma)$. We use the parameter γ_g to weight two kinds of distances: the Euclidean distance and the distances computed based on image gradient. For quick computation, the fast marching algorithm is used [10] to compute the geodesic distances. Then, the saliency model related to boundary prior is $S_i^{Bd} = g(i)$.

Similar to [8], the nonlinear combination of color contrast, color distribution and boundary prior S^{cmb} is defined by eq. (5),

$$S_i^{cmb} = S_i^{con} \times S_i^{dist} \times S_i^{Bd}. \tag{5}$$

The initial salient object region extracted based on saliency map is defined as $INTR = \{i | S_i^{cmb} \geq \eta\}$, the background region is $INTB = \{i | S_i^{cmb} < \eta\}$. The adaptive threshold $\eta = 1.5 \times S_{mean}$ where S_{mean} is the mean saliency over the entire saliency map.

3 Classifier Composition for Automatic Labeling

3.1 Classifier Based on Color Information

The basic features for pixel p is RGB color and the feature vector is represented as $I_p = RGB_p$. We define FG to represent the salient object and BG to represent the background. The color information contained in the sets $INTR$ and $INTB$ are modeled as Gaussian mixture model (GMM), respectively. Let the color

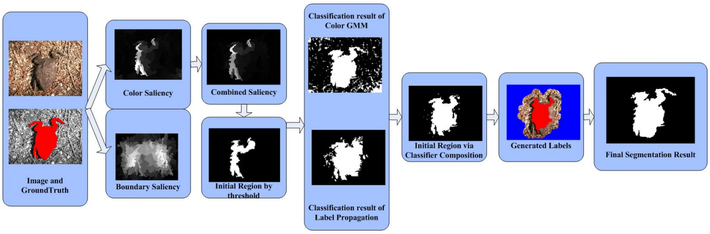


Image and GroundTruth
Color Saliency
Boundary Saliency
Combined Saliency
Initial Region by threshold
Classification result of Color GMM
Classification result of Label Propagation
Initial Region via Classifier Composition
Generated Labels
Final Segmentation Result

Fig. 1. The flow-chart of automatic seed generation. The pixels marked as red are the labels for objects and pixels marked as blue are background labels.

models be represented by GMM $\{\alpha_c, \mu_c, \Sigma_c\}_{c=1}^{C}$ in the RGB color space, where $\alpha_c, \mu_c, \Sigma_c$ represent the set of weight, mean color and covariance matrix of the $c-$th component, respectively. The mixture distribution of I_x can be formulated as a linear superposition of Gaussians in the form:

$$V(I_x|l) = \sum\nolimits_c \alpha_{cl} N(I_x|\mu_{cl}, \Sigma_{cl}), l \in \{FG, BG\}, \tag{6}$$

where $\{\alpha_{cl}, \mu_{cl}, \Sigma_{cl}\}$ represent the weight, the mean color and the covariance matrix of the $c-$th component learned from color information of class l, $l \in \{FG, BG\}$. In our experiments, GMM with 5 components are used to represent the color models in each class. Then, the posterior probability at each pixel p of the image is:

$$\mathrm{P}_{gmm}(F_p = l|I_p) = \frac{V(I_p|l)}{V(I_p|FG) + V(I_p|BG)}, l \in \{FG, BG\}. \tag{7}$$

The basic classifier is a function mapping the image space to figure-ground classification space:

$$H_{col}(p(I_i; F_i)) = \begin{cases} 1, p(F_i = FG|I_i) > p(F_i = BG|I_i) \\ 0, p(F_i = FG|I_i) \leq p(F_i = BG|I_i), \end{cases} \tag{8}$$

where $p(F_i|I_i)$ is the posterior probability associated with label F_i at pixel i. F_i is the label at pixel i and $F_i \in \{FG, BG\}$.

3.2 Classifier Based on Label Propagation

Given a point set $X = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$ and a label set $L = \{1, ..., c\}$. The indication vector is $y = \{y_1, ..., y_n\}^T$, in which $y_i = 1$ if x_i is labeled as salient object, and $y_i = 0$ otherwise. We set $y_i = 1$ for pixel $i \in INTR$. Let $f : X \rightarrow R^n$ represent a propagation function which assigns a value f_i to each point x_i. A graph $G = (V, E)$ is built on the points set. The edges E are weighed by an affinity matrix $W = [w_{ij}]_{n \times n}$. Given the graph, the degree matrix is

$D = diag\{D_{11}, ..., D_{nn}\}$, where $D_{ii} = \sum\limits_{j} w_{ij}$. Similar to [11], the optimization label propagation problem is:

$$\min_{\mathbf{f}:f(x)\in\mathbb{R}} \quad Q(\mathbf{f}) = \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\omega_{kj}\left(\frac{1}{\sqrt{D_{kk}}}f_k - \frac{1}{\sqrt{D_{jj}}}f_j\right)^2 + \theta\sum_{k=1}^{n}\left(f_k - y_k\right)^2, \quad (9)$$

where θ controls the balance between the smoothness constraint and fitting constraint. The result function with unnormalized Laplacian matrix is:

$$f^* = (D - \alpha W)^{-1}y, \quad (10)$$

where $\alpha = \frac{1}{1+\theta}$. f^* can be also interpreted as a probability and we define $P_{lp} = f^*$. Then, the classifier related to label propagation is described as:

$$H_{lab}(f(I_i)) = \begin{cases} 1, f^*(I_i) > \tau \\ 0, f^*(I_i) \leq \tau, \end{cases} \quad (11)$$

where τ is the adaptive threshold and we set $\tau = 1.5 \times \frac{\sum f^*(I_i)}{n}$.

3.3 Automatic Labeling

To divide the image region into several regions via Classifier Composition. Two basic pixel sets $A = \{i|H_{col}(i) > 0\}$ and $B = \{i|H_{lab}(i) > 0\}$ are defined. $\bar{A}$ and $\bar{B}$ are the related complements.

$$\begin{array}{l} C1 = \{i|i \in A \cap B\}, C2 = \{i|i \in A \cap \bar{B}\}, \\ C3 = \{i|i \in B \cap \bar{A}\}, C4 = \{i|i \in \bar{B} \cap \bar{A}\}. \end{array} \quad (12)$$

We use pixels in $C1$ to generate the foreground seeds LF. The pixels with low saliency value is contained in set $SAL = \{i|S^{cmb}(i) < 0.1\}$. The pixel in set $D = C4 \cup SAL$ is utilized to generate the background seeds. We shrink the initial region $C1$ to avoid inexact boundaries and form an accurate object labels. By shrinking pixels in set D, a ring region is obtained which are taken as the background labels LB. The process of automatical labeling is illustrated in Fig. 1.

4 Formulation of Salient Segmentation

Image segmentation can be modeled with a conditional random field (CRF). Consider a random field F defined over a set of variables $\{F_1, F_2, ..., F_N\}$. The domain of each variable is a set of labels $L = \{\ell_1, \ell_2, ..., \ell_k\}$. Let $I = \{I_1, .., I_N\}$ be the observed data corresponding to image information and N is the image dimension. I_i is the feature vector at pixel i and $I_i = \{RGB_i, LAB_i, S_i^{cmb}\}$. F_i represents the label assigned to pixel i. In our model, two features, color model P_{gmm} and label propagation probability P_{lp} are used. Let w be an $N \times 2$ matrix and $w = \{w_1, w_2\}$, where $w_i = [w_{i1}, w_{i2}, ..., w_{iN}]^T$ is an N-dimensional vector

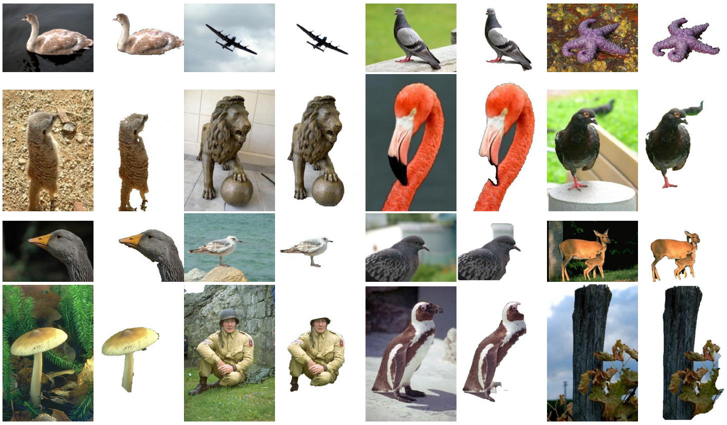


Fig. 2. Segmentation results of our method. The corresponding results are listed next to the original images.

related to a feature. We formulate the segmentation problem as a binary labeling task and the energy function $E(F|I,w)$ takes the form:

$$E(F|I,w) = w_1^T E_1^{Col}(F|I) + w_2^T E_1^{Lp}(F|I) + \tau E^{Pair}(F|I). \tag{13}$$

The energy function $E_1^{Col}(F|I)$ is the energy related to color information and $E_1^{Lp}(F|I)$ represents the single-site clique potentials related to label propagation probability. $E^{Pair}(F|I)$ is the pair-site clique potentials and the parameter τ is a control weight for the pairwise constraint. We set $\tau = 1$ in the experiments. In the process of CRF construction, labels LF, LB (described in section 3.3) are used to compute P_{gmm} and P_{lp}. Then, the common unary potential for two features is:

$$\begin{array}{l} E_1^{Col/Lp}(F|I) = \{V_1^{Col/Lp}(F_1), ..., V_N^{Col/Lp}(F_N)\}^T, \\ {V_1}^{Col/lp}(F_q = l) = -\log(\mathrm{P}_{gmm/lp}(I_q; F_q = l)). \end{array} \tag{14}$$

In the experiments, we set $w_i = \{0.5, 0.5\}^T$. The pairwise term between neighbor nodes is computed based on the low-level features (such as RGB color, LAB color and saliency value). The related pixel pairwise term is defined as:

$$E^{pair}(i,j) = exp(-|I_i - I_j|/2\sigma^2), i,j \in NEB, \tag{15}$$

where NEB is set of pixels in neighborhood and $\sigma = 0.5$ for the experiments.

5 Performance Evaluation

In this section, we evaluate the performance of our method on Berkeley [15] and Weizmann [13] databases. Some segmentation results are illustrated in Fig. 2.

Table 1. Performance comparison of our method with other segmentation methods: F-measures of our method and 4 state-of-the-art segmentation algorithms by evaluating them on the Weizmann single object database.

Methods	**F-measure(%)**	**Remarks**
[6] With auto-context cues	0.91±0.013	Automatic
Proposed Framework	0.89±0.002	Automatic
[6] Without auto-context cues	0.88±0.011	Automatic
[12] Unified approach	0.87±0.011	interactive
[13] Cues integration	0.86±0.012	Automatic
GMM+Initial	0.86±0.011	Automatic
Label Propagation+Initial	0.85±0.012	Automatic
[14] Texture segmentation	0.83±0.016	Automatic

Empirical results show that our method can extract salient object efficiently, and is able to deal with the images with weak boundary or complex background. The F-measure score ($F = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$) is computed as well for objective comparison on Weizmann single object segmentation database, and the result is listed in Table 1. We compare our method with four state-of-the-art methods and the F-measure scores of these methods are quoted from [6]. Based on the initial seeds (described in Section 3.3), the scores of segmentation results using CRF (described in Section 4), by GMM classifier in eq. (8) and by Label Propagation classifier in eq. (11) are presented in Table 1 as well. It is noticed that the F-measure score of our method (using CRF) outperforms all the baselines except for the result of [6], which applies the context cues. The performance of our method can be further improved by integrating more discriminative features or refined iteratively.

6 Conclusion

In this paper, we have proposed a framework to extract salient objects from images automatically. Firstly, we propose a saliency model to estimate the initial object region exploiting the low level color features and prior information. Secondly, the seeds are generated automatically by classifiers composition to obtain more precise initial region. Finally, a CRF model is constructed to extract the salient objects. Experimental results show that the proposed method can achieve better performance than baselines on some popular segmentation benchmarks. In future work, we will explore how to incorporate high-level classifiers into the proposed segmentation model.

Acknowledgments. This research is partly supported by NSFC, China (No: 61273258, 31100672)Ph.D. Programs Foundation of Ministry of Education of China (No.20120073110018).

References

1. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In: 2001 Eighth IEEE International Conference on Computer Vision (ICCV), vol. 1, pp. 105–112. IEEE (2001)
2. Achanta, R., Susstrunk, S.: Saliency detection using maximum symmetric surround. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 2653–2656. IEEE (2010)
3. Rahtu, E., Kannala, J., Salo, M., Heikkilä, J.: Segmenting salient objects from images and videos. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 366–379. Springer, Heidelberg (2010)
4. Jung, C., Kim, B., Kim, C.: Automatic segmentation of salient objects using iterative reversible graph cut. In: 2010 IEEE International Conference on Multimedia and Expo (ICME), pp. 590–595. IEEE (2010)
5. Liu, Z., Shi, R., Shen, L., Xue, Y., Ngan, K.N., Zhang, Z.: Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut. IEEE Transactions on Multimedia 14(4), 1275–1289 (2012)
6. Xue, J., Wang, L., Zheng, N., Hua, G.: Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting (2013), http://www.sciencedirect.com/science/article/pii/S0031320313001581
7. Park, K.T., Moon, Y.S.: Automatic extraction of salient objects using feature maps. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, p. I–617. IEEE (2007)
8. Fu, K., Gong, C., Yang, J., Zhou, Y.: Salient object detection via color contrast and color distribution. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 111–122. Springer, Heidelberg (2013)
9. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 29–42. Springer, Heidelberg (2012)
10. Sethian, J.A.: Fast marching methods. SIAM Review 41(2), 199–235 (1999)
11. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. Advances in Neural Information Processing Systems 16(753760), 284 (2004)
12. Bagon, S., Boiman, O., Irani, M.: What is a good image segment? A unified approach to segment extraction. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 30–44. Springer, Heidelberg (2008)
13. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE (2007)
14. Galun, M., Sharon, E., Basri, R., Brandt, A.: Texture segmentation by multiscale aggregation of filter responses and shape elements. In: Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV), pp. 716–723. IEEE (2003)
15. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(5), 530–549 (2004)

ICA for Separation of Respiratory Motion and Heart Motion from Chest Surface Motion

Ghufran Shafiq, Yubo Wang, Sivanagaraja Tatinati, and Kalyana C. Veluvolu

School of Electronics Engineering, Kyungpook National University,
702-701 Daegu, Republic of Korea
ghufran.shafiq@knu.ac.kr

Abstract. Chest surface movement contains information of respiration and heart activity which are considered vital parameters. However, it is important to separate respiratory and cardiac information in order to perform further analysis. For this purpose, Independent Component Analysis (ICA) was applied to multiple simultaneously recorded chest surface movement signals. Successful separation of cardiac pattern is demonstrated and compared with ECG. This methodology can be used to further develop non-obtrusive ways to monitor vital physiological parameters in the form of wearable sensors.

Keywords: Cardiography, Respiration motion, ICA, vital parameters.

1 Introduction

Acquisition and analysis of physiological signals have gained high importance in pursuing improved health care systems. The physiological signals generated due to respiratory and cardiac systems are considered vital [6]. The existing clinical modalities to monitor such signals restrict the monitoring to the hospital settings and owing to the obtrusiveness resulting from attaching number of electrodes to the patient's body, long-term monitoring doesn't seem a favorable option. Alternatively, non-obtrusive ways are being developed which utilize the quasi-periodic movement of the surface chest wall. It has been reported that this movement, ranging from 4-12 mm [1], is proportional to the lung volume changes during respiration [2]. In addition, a smaller quasi-periodic chest wall movement due to beating of heart is reported ranging from 0.2-0.5mm [3].

The aim of this study is to separate the respiratory and cardiac components from the surface chest motion that can be potentially useful for monitoring apnea, stress, depression, anxiety and heart diseases etc. Several works have tried to separate the cardiac source component from the surface chest motion. In [4], the cardiac source component is identified from microwave doppler radar in the absence of respiration. [6] used doppler radar for obtaining the surface chest wall motion. This motion is also recorded with optical interferometer in [5,2]. Apart from acquiring the physiological signals, various approaches have been adopted for extracting the cardiac component from the composite chest movement signal. For instance, a simplistic approach presented in [2] is based

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 592–599, 2013.

on two band-pass filters for separating out cardiac and respiratory components. However, the narrow bandwidth of these filters results in lack of details in the extracted pattern. Further, the methods based on band-pass filters cannot be generalized for all subjects. Wavelet decomposition [6] was used to separate the cardiac component from the surface chest motion. However, in these works, the extracted pattern from free breathing lacks the fine details as compared to the patterns presented for breath-hold recordings. Recently, in [5] a high detailed cardiac component pattern is extracted that shows the synchronization with the corresponding simultaneous ECG recording. But results were demonstrated for breath-hold maneuver only.

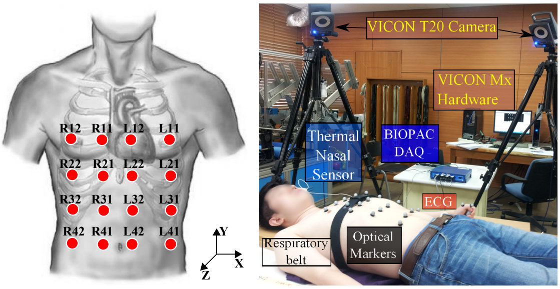
Our goal was to separate the detailed cardiac pattern under normal breathing conditions in attempt to open new doors for study and analysis of these patterns. For this purpose, we adopted a Blind Source Separation (BSS) approach since the recorded physiological signal under consideration i.e. surface chest motion is a mixture of mainly respiratory source and cardiac source, while no apriori information about these sources is needed. Application of ICA on multiple simultaneously recorded surface chest motion can separate the cardiac source component which can be generalized for subjects, higher details of pattern for sufficient length can be obtained.

2 Methods and Materials

2.1 Experimental Setup and Protocol

Experimental verification of the proposed approach was performed on the data acquired from healthy subjects. In the preliminary phase of investigation, 5 healthy male subjects aging in the range of 22-27 years were included in the study. Written consent of all the subjects was taken prior to the experiment.

The data acquisition involved recording of surface chest wall movement, Electrocardiogram (ECG) and referential respiratory signals. Complete experimental setup is shown in Fig. 1(b).



(a) (b)

Fig. 1. (a) Placement of markers and Labeling scheme, (b) Experimental Setup

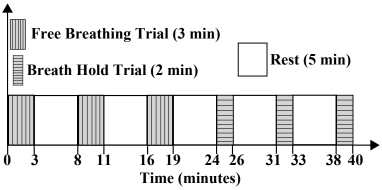


Fig. 2. Timing Diagram for the Recording Session

Complete Block Diagram of data acquisition and our approach is illustrated in Fig. 3. The surface chest wall movement was acquired by six VICON infrared camera system (Vicon Motion Systems Ltd., UK), that could track the movement of retro-reflective optical markers. These markers were placed on the anterior upper body in a 4×4 grid as shown in Fig. 1(a). The placement of these markers was done to cover the motion of full trunk and to attain the maximum possible chest wall movement due to cardiac source (near xiphoid process). On the other hand, the cameras were placed in near-circular pattern around the subject to ensure that each marker placed was visible by at least three cameras for reliable 3D reconstruction of motion. The sampling rate for the VICON camera system was set to 100 frames per second. The ECG and the reference respiratory signals (respiration belt and thermal nasal sensor) were recorded by BIOPAC biological data acquisition system (BIOPAC Systems Inc., USA). The sampling frequency for the BIOPAC recorder was set to 500 Hz.

All recordings were made with subject in supine position. Six trials were recorded for each subject out of which three trials were recorded under normal breathing, while in last three trials, subject was asked to perform breath-hold maneuver as shown in Fig. 2. Each breath-hold trial contains multiple breath-hold maneuver attempts. However, the duration for such attempts were dependent on subject's discretion and comfort.

Since we are more interested in the motion perpendicular to the frontal plane of chest, only z-axis data is retained for further processing (refer to Fig. 1(a)). Further preprocessing steps include downsampling of displacement signals to 20 Hz, bandpass filtering from 0.1-4 Hz to improve SNR, input selection (explained in results) and making the signals zero-mean and unit-variance to fulfil conditions for ICA. For input selection, 8 markers data were selected out of total 16 markers data. The selection of markers were made such that it included 2^{nd} row, where one marker is intended to cover the precordial area, as well as 4^{th} row placed on abdomen (see Fig. 1(a) and Fig. 4). The selected preprocessed displacement signals are then fed as input to ICA algorithm and as a result, cardiac source component is obtained as one of the Independent component. The referential respiratory signals are kept for future analysis.

2.2 Independent Component Analysis

The chest surface motion recorded with multiple marker contains different magnitude of cardiac component. To separate the underlying respiratory and cardiac

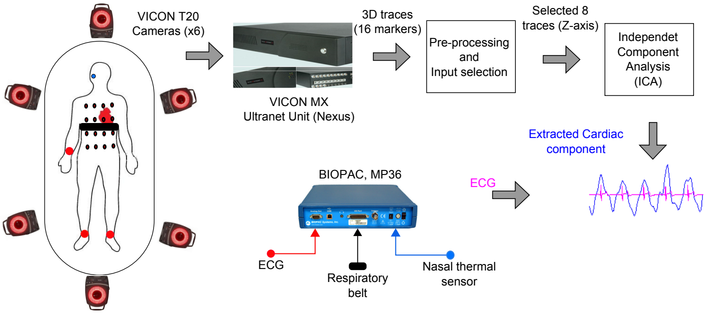


Fig. 3. Illustration of Data acquisition and Methodology

sources, Independent Component Analysis (ICA) is employed which was originally designed in [7] to perform BSS. Let $\mathbf{x} = \{x_1, x_2, ..., x_8\}^T$ be the collection of motion recorded from 8 selected markers and $\mathbf{s} = \{s_1, s_2, ..., s_8\}^T$ be the 8 underlying sources of chest surface motion (equal observations and sources are assumed for simplicity), then these signals can be expressed in terms of ICA model as $\mathbf{x} = A\mathbf{s}$, where A is the 8×8 mixing matrix that corresponds to the weights of the underlying source signals in the observed chest surface signals.

The goal of ICA is to obtain the collection of sources $\hat{\mathbf{s}}$ based on the estimation of a demixing matrix W such that $\hat{\mathbf{s}} = W\mathbf{x}$, where $\hat{\mathbf{s}} = \{\hat{s}_1, \hat{s}_2, ..., \hat{s}_8\}^T$ represents the collection of 8 estimated source signals, while W is the 8×8 demixing matrix. One approach to this problem is to find W such that resulting sources are maximally non-gaussian. For this purpose, negentropy based measure is employed in FastICA [8]. Owing to its fast convergence and ability to separate sub-gaussian components [9], FastICA algorithm is employed to estimate the underlying sources as the desired sources are quasi-periodic having sub-gaussian distributions. The 8 independent components obtained from ICA contains the desired cardiac source component which can be identified by examining the power spectral density (PSD) plots of all the components.

It is assumed that the sources under consideration do not have significant spatial movement, therefore mixture is linear. Whereas, the condition of statistical independence does not need to be exactly true in practice [10]. Therefore, ICA model holds for the given observed chest wall movement signals.

3 Results

Application of ICA on selected displacement signals for Subject # 1 and resulting components are shown in Fig. 4. Fig. 4(a) z-axis displacement for the selected 8 markers for 20 seconds; Fig. 4(b) shows PSD of L22 marker as it is placed near the xiphoid process and observable cardiac source component is expected. Application of ICA on selected inputs (see Methods) yielded 8 ICs. The fundamental band from PSD of IC8 and ECG are having same bandwidth (0.24) Hz with peak frequency of 0.91Hz i.e. 54.6 beats per minute.

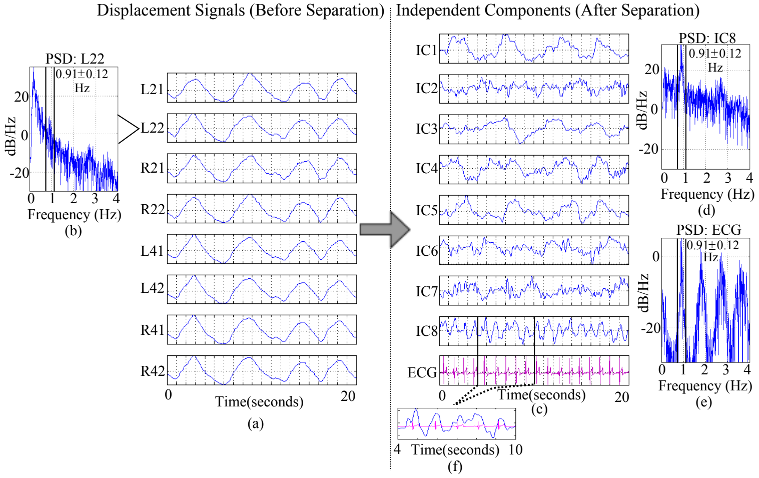


Fig. 4. Subject # 1: Before and after source separation: (a) Displacement Signals before separation, (b) PSD of one marker, (c) Components after separation, (d) PSD of cardiac component, (e) PSD of ECG, (f) 4-10 second preview of IC 8 with ECG

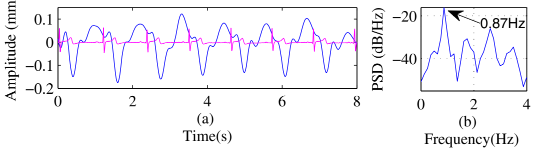


Fig. 5. Subject # 1: (a) Cardiac source pattern, (b) PSD under breath-hold maneuver

During breath-hold trials, respiration is suppressed and a detailed cardiac pattern in the chest motion can be observed. 8 second breath-hold chest wall movement obtained from L22 is presented along with the corresponding ECG for subject # 1 in Fig. 5. This pattern is very similar to the ones presented in [4] and [5]. The base-line drift was removed using a zero-phase FIR filter. The observed pattern shows a good accordance with the corresponding ECG. The corresponding PSD is shown in Fig. 5(b). The dominant peak centered at 0.87 Hz indicates the average Heart Rate of 52.2 bpm, whereas, the next three peaks are the 2^{nd}, 3^{rd} and 4^{th} harmonics.

A comparison between the PSD of separated cardiac component and the corresponding ECG is illustrated for two subjects in Fig. 6. The top row in the figure corresponds to subject # 2 (BMI = $20.7 kg/m^2$), while the bottom row in the figure corresponds to subject # 3 (BMI = $22.3 kg/m^2$). Since ECG and the separated cardiac component belong to two different domains (electrical and mechanical), therefore, we expect only the fundamental frequency bands to be similar in both the PSDs. For this reason and to ensure the convenient visibility,

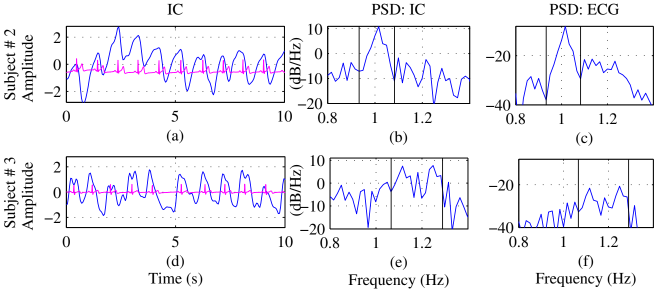


Fig. 6. (a,d): Separated cardiac component for two subjects along with corresponding ECGs, (b,e): PSDs of separated components, (c,f): PSDs of ECG on the right panel

the frequency axis is limited from 0.8 – 1.4 Hz. For subject # 2, the fundamental band of the separated cardiac component was identified from 0.933 – 1.083 Hz (peaking at 1.0166 Hz) and is highlighted with two vertical lines. The corresponding PSD of ECG to the right also indicates the same bandwidth as indicated by vertical lines bounding the same range of frequencies (and same dominant frequency). Similarly, for subject # 3, the fundamental bands indicated by vertical lines at 1.066 Hz and 1.29 Hz for both the PSDs lie in the same frequency range (with peak frequency of 1.249 Hz for IC and 1.25 Hz for ECG). The alignment of separated component with ECG along with the same fundamental frequency bands suggests that the separated component is indeed generated due to mechanical activity of heart.

Statistical comparison of the power in the fundamental respiratory and cardiac band before and after the separation for all 5 subjects in all trials is shown in Fig. 7. Except for subject 4, the average power for cardiac source after separation (SC) is higher than the respiratory source after the source separation (SR). Even

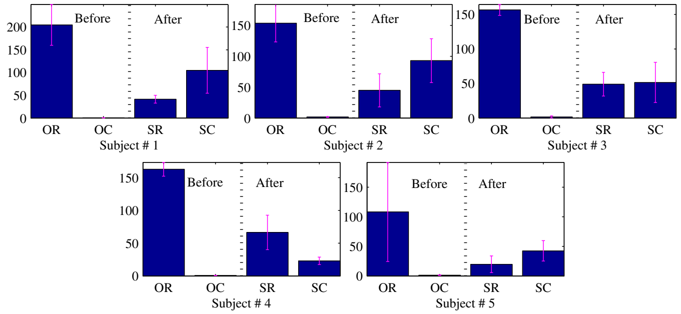


Fig. 7. Avg. Power in fundamental Respiratory and Cardiac Bands. OR and OC represents power respiratory and cardiac power respectively before separation; SR and SC represents respiratory and cardiac power respectively after separation

though, the power of cardiac source is still boosted for all the subjects after the cardiac source separation using the proposed approach.

4 Conclusion and Discussion

The application of ICA to multiple simultaneously recorded chest surface motion yields detailed cardiac source pattern which consistently aligns with the corresponding ECG signal in most of the cases. Moreover, the similarity of PSD found in fundamental bands of ECG and separated cardiac source indicates accurate Heart beat calculation. Even when extracted component is dominant by respiratory source, the cardiac source power is significantly boosted as compared to the power before source separation. The proposed application is advantageous over the previously used methods to separate the cardiac source component since it is generalized for all the subjects as opposed to band-pass filters and wavelet decomposition (finding optimal wavelet and appropriate decomposition level).

The final goal for this approach is to employ light-weight accelerometers in the form of wearable sensors to obtain chest surface displacement signal for classification of respiratory and cardiac sources. In this way, non-obtrusive monitoring of detailed vital physiological signals can be made possible without interfering one's daily activities or need of any special assistance.

Acknowledgments. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant No. 2011-0023999).

References

1. De Groote, A., Wantier, M., Cheron, G., Estenne, M., Paiva, M.: Chest wall motion during tidal breathing. Journal of Applied Physiology 83(5), 1531–1537 (1997)
2. Silva, A.F., Carmo, J.P., Mendes, P.M., Correia, J.H.: Simultaneous cardiac and respiratory frequency measurement based on a single fiber Bragg grating sensor. Measurement Science and Technology 22(7), 75801 (2011)
3. Ramachandran, G., Singh, M.: Three-dimensional reconstruction of cardiac displacement patterns on the chest wall during the P, QRS and T-segments of the ECG by laser speckle inteferometry. Medical and Biological Engineering and Computing 27(5), 525–530 (1989)
4. Lin, J.C., Kiernicki, J., Kiernicki, M., Wollschlaeger, P.B.: Microwave apexcardiography. IEEE Transactions on Microwave Theory and Techniques 27(6), 618–620 (1979)
5. Obeid, D., Zaharia, G., Sadek, S., El Zein, G.: Microwave doppler radar for heartbeat detection vs electrocardiogram. Microwave and Optical Technology Letters 54(11), 2610–2617 (2012)

6. Mikhelson, I.V., Bakhtiari, S., Elmer, T.W., Sahakian, A.V.: Remote Sensing of Heart Rate and Patterns of Respiration on a Stationary Subject Using 94-GHz Millimeter-Wave Interferometry. IEEE Transactions on Biomedical Engineering 58(6), 1671–1677 (2011)
7. Comon, P.: Independent component analysis, A new concept. Signal Processing 36(3), 287–314 (1994)
8. Hyvarinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Computation 9(7), 1483–1492 (1997)
9. Hyvarinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10(3), 626–634 (1999)
10. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons (2004)

Real-Time Head Detection with Kinect for Driving Fatigue Detection

Yang Cao[1] and Bao-Liang Lu[1,2]

[1] Center for Brain-like Computing and Machine Intelligence
Department of Computer Science and Engineering
[2] MOE-Microsoft Key Lab. for Intelligent Computing and Intelligent Systems
Shanghai Jiao Tong University
800 Dongchuan Road, Shanghai 200240, P.R. China

Abstract. Nowadays, depth cameras such Microsoft Kinect make it easier and cheaper for us to capture depth images. It becomes practical to use depth images for detection in consumer-grade products. In this paper, we propose a novel and simple real-time method to detect human head in depth image for our driving fatigue detection system, based on the elliptical shape of human head. Experiments show that our method can successfully detect human head in different light conditions and across different head poses. We integrate this detection algorithm into our driving fatigue detection system, and see remarkable improvements both in detection rate and detection speed.

Keywords: Head Detection, Depth Image, Kinect, Fatigue Detection.

1 Introduction

Human head detection is often the fundamental step in many computer vision applications, such as head or face tracking, face recognition, face expression analysis and gender classification. Given a color image, an infra-red image, a depth image or a combination of them, the goal of head detection is to find the locations and sizes of all human heads in the image. Head detection is difficult if different light conditions and different head poses are taken into consideration.

Traditionally, human head detection is accomplished on color images. Since human face has the most significant features in head, we often do face detection as a way of head detection. Face detection has been studied for decades. In [2] face detection methods before year 2000 are nicely surveyed. In the 2000's, Viola and Jones made a great contribution to this field by their excellent work in [1]. Many later works are based on or inspired by the approach in [1]. Works after year 2000 are surveyed in [3].

Human head detection on depth image has also been studied for a long time, which can be broadly classified into two types. One type of head detection methods on depth images tries to detect features of faces such as nose tips, eyes and cheeks. For example, Colombo *et al.* [4] proposed to detect human face on depth image through an analysis of the curvature of face. Chew *et al.* [5] proposed to

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 600–607, 2013.

detect nose tips in a depth image by calculating effective energy. For this type of head detection methods, the advantage is that organs on faces are detected directly, which provides us more information such as locations of each organ instead of just head location and size. The disadvantage is that detection rate often decreases dramatically for depth image with moderate noise. The other type of head detection methods on depth images tries to detect human heads using general information such as the elliptical shape of head. Xia *et al.* [7] used 2D chamfer distance matching to find candidate head locations according to the contour of head and shoulder, and then used a hemisphere to fit the head. Suau *et al.* [8] first extracted foreground pixels in the depth image, and then used a binary elliptical template to search for the head location. This type of head detection methods can tolerate moderate noise in the depth image, and is often more robust to different head poses.

Our driving fatigue detection system operates according to the facial expression of the driver, so accurate and fast head detection is an important step. Generally speaking, color image can provide more information than depth image, but is more sensitive to light condition. Considering the complex light condition during driving, the head detection method proposed in this paper utilizes depth information only. We ignore facial details in the depth image, and use the elliptical shape of head as a clue. Methods proposed in [7,8] also use the same clue to detect human head. The difference between their methods and our method is that, their methods both include a matching stage, *i.e.* shifting a template in the depth image pixel by pixel to find a best match, which greatly slow down the speed of head detection. For example, the method in [8] can only do real-time tracking at resolution 160 × 120. However, combined with a tracking trick, our method can achieve real-time tracking at resolution 640 × 480.

Our method contains three key steps: depth image split, contour extraction and ellipse fitting. Section 2 describes our method in detail. Section 3 evaluates our method on the Kinect face database proposed in [6]. Section 4 shows the integration of our driving fatigue detection system and the method proposed in this paper, followed by the conclusion and discussions in section 5.

2 Method

2.1 Assumptions

In our head detection method, we make two heuristic assumptions about human head.

a) Approximately, human head has an elliptical shape.
b) The depth values of human head are continuous.

Strictly speaking, these two assumptions doesn't hold for every people every time. For example, if a person has very bushy hair or wears accessories such as caps, his/her head may not have an elliptical shape in depth image. However, we find these two assumptions do hold in most cases. In fact, both [7] and [8] make assumptions which are similar with ours.

2.2 Work-Flow

Fig. 1 illustrates the overall process of our head detection method.

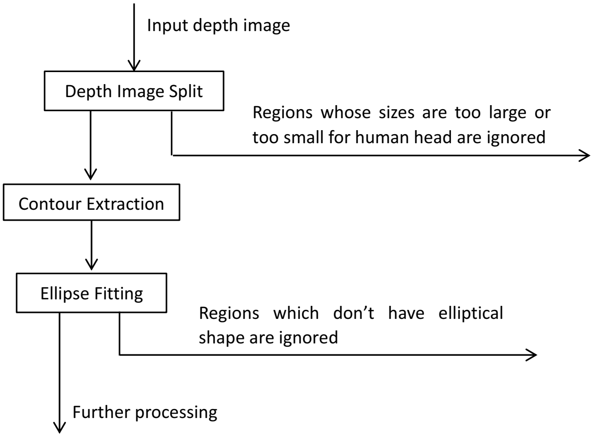

Input depth image
Depth Image Split
Regions whose sizes are too large or too small for human head are ignored
Contour Extraction
Ellipse Fitting
Regions which don't have elliptical shape are ignored
Further processing

Fig. 1. The work-flow of our head detection method

First, we split the depth image into regions according to the depth value of each pixel. Because we assume the depth values of human head are continuous, we can set a threshold and split depth image at locations where the change of depth is larger than the threshold. Regions whose size is much larger or much smaller than normal human head will be ignored. Second, we extract the contour of each region. For the extraction step, we use an algorithm which 'walks' along the contour of each region. Third, after we get the contour of each region, an ellipse fitting algorithm will be used and the similarity between the region and an ellipse will be calculated. Regions with an elliptical shape will be returned for further processing.

2.3 Depth Image Split

Image split is a kind of image segmentation. According to [9] there is quite a number of image split algorithms. Considering effectiveness, simplicity and speed, we choose image split based on computing connected components which is enough for our method.

In our design we use breadth first search to compute connected components. Neighboring pixels whose difference of depth is less than threshold will be treated as connected pixels. Fig. 2 is an example of image split. Here different regions are colored with different colors. We can see that the region of head has been split out. Notice that regions much larger or much smaller than normal human head will be removed.

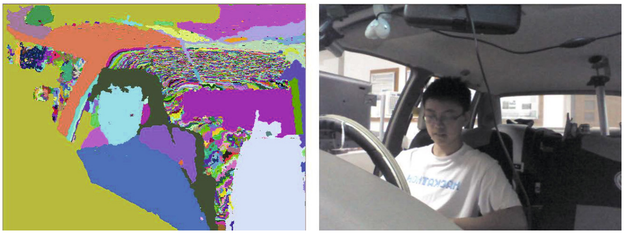


Fig. 2. An example of image split result (The right color image is for reference)

2.4 Contour Extraction

After the depth image is split into different regions, contour extraction is performed. Contours are expressed by contour points. Fig. 3 gives an example of a region (left figure) and it's corresponding contour points (right figure). One pixel in depth image is represented by one block in Fig. 3.

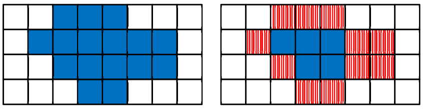


Fig. 3. A region in depth image and it's corresponding contour points

We use an contour extraction algorithm which 'walks' along the boundary of each region.

To represent contour points, a corresponding format is defined. This format contains not only the location of the point (*i.e.* the coordinates), but also the normal direction (*i.e.* the direction pointing to the 'outside' of the region). For example, if we express contour points in Fig. 3 using this format, we will have a list like this (starting from the top-left point)

$\langle 2,0,\mathrm{UP}\rangle\langle 2,0,\mathrm{LEFT}\rangle\langle 1,1,\mathrm{UP}\rangle\langle 1,1,\mathrm{LEFT}\rangle\langle 1,1,\mathrm{DOWN}\rangle\langle 2,2,\mathrm{LEFT}\rangle\ldots$

Each contour point is represented by three elements. The first two elements are the coordinates of the contour point, and the third element shows the normal direction of the contour point. Using this format, we can easily find and express the contour of a region.

The 'walking' procedure is a procedure of state transition, *i.e.* transition from a contour point to another contour point. The algorithm starts at a contour point of the region (top-left point of the region in our system), and 'walks' along the contour as the state transiting. When we get back to the start contour point, we know that we have found all the contour points of this region.

State transition depends on current contour point. We have four normal directions (UP, DOWN, LEFT, RIGHT), each direction has 3 situations, so we have 12 situations to consider in total. Fig. 4 gives us an example of the UP situation. DOWN, LEFT, RIGHT are very similar.

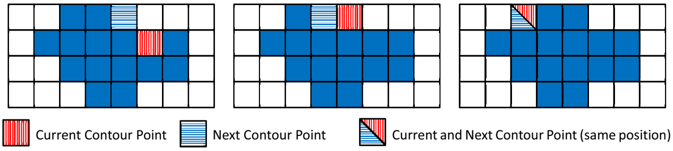


Fig. 4. Three cases of finding next contour point if current normal direction is UP

Notice that a certain location (coordinates) may contain more than one contour point. Fig. 5 is the contour extraction result of Fig. 2.

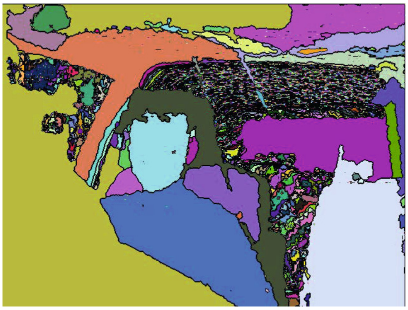


Fig. 5. An example of contour extraction result

2.5 Ellipse Fitting

Having found the contour of each region, we can use any ellipse fitting algorithm to fit an ellipse for the contour points. In our system, we choose the algorithm in [10]. We calculate fitness for each region using the formula below (less fitness value means higher fitness degree):

$$\frac{\sum_{i=1}^{n} \delta_i}{n \cdot h} \tag{1}$$

Here n is the number of contour points of the region, δ_i is the offset of the contour point to its corresponding point in the fitted ellipse (draw a line cross ellipse center and the contour point, the nearer intersection with the ellipse), and $h = \max\{\text{height of region}, \text{width of region}\}$. In conclusion, we calculate the normalized average offset of every contour points.

After we get the fitness of each region. We return the region whose fitness is less than a certain threshold as the human head we detected.

3 Experiments

We test our head detection method on a public available face database provided by Hg *et al.* [6]. In this database there are 31 persons, and each person has 17 poses (sitting in front of Kinect and looking at different positions with different facial expressions). For each pose, color image and depth image are taken at the same time for three times, so there are $31 \times 17 \times 3 = 1581$ color images and 1581 depth images. Both color images and depth images are taken using Microsoft Kinect, with a resolution 1280×960 for color images and a resolution 640×480 for depth images.

The test result is shown in Table 1. The detection rate is satisfying for most persons, and is not sensitive to head pose. But for persons with very bushy hair or beard, the detection rate drops heavily (boldface items in Table 1).

Table 1. Detection rate for each person in the database

Id	Detection Rate	Id	Detection Rate	Id	Detection Rate	Id	Detection Rate
1	88.24%	9	88.24%	**17**	**23.53%**	25	98.04%
2	90.20%	10	84.31%	18	100.00%	26	100.00%
3	100.00%	11	100.00%	**19**	**7.84%**	27	100.00%
4	100.00%	12	84.31%	20	100.00%	28	100.00%
5	64.71%	13	100.00%	21	100.00%	29	90.20%
6	**0.00%**	14	70.59%	22	100.00%	30	100.00%
7	100.00%	15	92.16%	23	94.12%	31	100.00%
8	**9.80%**	16	88.24%	24	100.00%	Average	83.05%

For detection speed, it takes about 170ms in average to do detection for an depth image of resolution 640×480 in my computer[1]. Compared with method proposed in [8], our method is about 2 times faster. The reason is that there are no template matching procedure in our method, which can greatly slow down the speed.

4 Integrate with Driving Fatigue Detection System

Fig. 6 shows the work-flow of our driving fatigue detection system. Here, head detection in depth image is an important step which speeds up the whole system. Originally, we detect human face in color image using Viola-Jones algorithm [1] directly, which requires about 1800ms for an image of resolution 1280×960. Now, we first detect human head in depth image, which requires about 170ms for an depth image of resolution 640×480, and then use traditional algorithm to detect face in the head region, which requires about 70ms.

[1] CPU: Pentium T4400 2.2GHz, RAM: 4GB

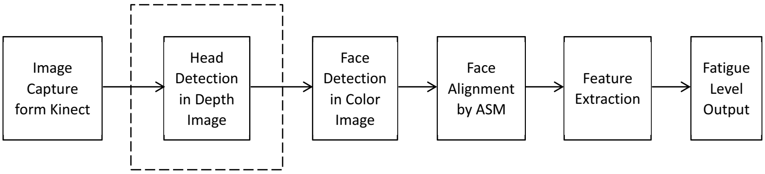


Image Capture form Kinect
Head Detection in Depth Image
Face Detection in Color Image
Face Alignment by ASM
Feature Extraction
Fatigue Level Output

Fig. 6. The work-flow of our driving fatigue detection system

After we get the accurate location of the driver's face, we use ASM (Active Shape Model) to determine the location of each organ on the face. This is done on color image. Then the appearance of each organ on the face can help us determine the facial expression of the driver. Currently, we use the shape of driver's mouth as a fatigue feature, which is expressed by a value:

$$d = \min\{100, \frac{h}{w} \times 100\} \tag{2}$$

Here h is the height of mouth, and w is the width of mouth. Therefore d increases as the driver opens his/her mouth. After we calculate d for each frame, we use formula below to calculate fatigue degree:

$$f = \min\{100, \frac{\sum_{i=1}^{n} d_i}{n} + 20t\} \tag{3}$$

Here n is the number of frames in 3 minutes, t is the number of yawns. For the calculation of t, we set a threshold for d. If d is larger than the threshold, we add 1 to t, and wait 10 seconds (we assume that yawn will not happen twice within 10 seconds or last for more than 10 seconds).

5 Conclusion and Future Work

In this paper, we proposed a novel human head detection method for depth image, which is both simple and robust. From the experiments, we can see that our method can achieve comparable detection rate and can reduce detection time.

This method utilizes depth information only, so it is not sensitive to light condition. The validity of the method is demonstrated by integrating the head detection method into our driving fatigue detection system. In the future, we plan to improve our method so that it can deal with very bushy hair and beard. For the driving fatigue detection system, we will add more features to get more accurate result.

Acknowledgements. This work was partially supported by the National Natural Science Foundation of China (Grant No. 61272248), the National Basic Research Program of China (Grant No. 2013CB329401) and the Science and Technology Commission of Shanghai Municipality (Grant No. 13511500200).

References

1. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
2. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34–58 (2002)
3. Zhang, C., Zhang, Z.: A survey of recent advances in face detection. Microsoft Research (2010)
4. Colombo, A., Cusano, C., Schettini, R.: 3D face detection using curvature analysis. Pattern Recognition 39(3), 444–455 (2006)
5. Chew, W.J., Seng, K.P., Ang, L.M.: Nose tip detection on a three-dimensional face range image invariant to head pose. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1 (2009)
6. Hg, R.I., Jasek, P., Rofidal, C., Nasrollahi, K., Moeslund, T.B., Tranchet, G.: An RGB-D Database Using Microsoft's Kinect for Windows for Face Detection. In: International Conference on Signal Image Technology and Internet Based Systems, pp. 42–46 (2012)
7. Xia, L., Chen, C.C., Aggarwal, J.K.: Human detection using depth information by Kinect. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 15–22 (2011)
8. Suau, X., Ruiz-Hidalgo, J., Casas, J.R.: Real-time head and hand tracking based on 2.5 D data. IEEE Transactions on Multimedia 14(3), 575–585 (2012)
9. Szeliski, R.: Computer vision: algorithms and applications. Springer (2010)
10. Fitzgibbon, A.W., Fisher, R.B.: A buyer's guide to conic fitting. In: Proceedings of the 6th British Conference on Machine Vision, vol. 2, pp. 513–522 (1995)

A Classification-Based Approach for Retake and Scene Detection in Rushes Video

Quang-Vinh Tran[1], Duy-Dinh Le[2,3], Duc Anh Duong[3], and Shin'ichi Satoh[2]

[1] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
tqvinh@fit.hcmus.edu.vn
[2] National Institute of Informatics, Tokyo, Japan
{ledduy,satoh}@nii.ac.jp
[3] Multimedia Communications Lab, University of Information Technology,
VNU-HCM, Ho Chi Minh City, Vietnam
ducda@uit.edu.vn

Abstract. Retake detection has been a challenging problem in rushes video summarization. Previous approaches represent video segments as a sequence of labels then find retakes by grouping similar sub-sequences using some sequence alignment algorithm. However, these kinds of representation usually lead to unsatisfactory results because it is difficult to know the number of labels needed for a video. In our method, instead of quantizing each video segment into a label, we formulate it as a binary classification problem between pairs of segments. We use this information as the input for the Smith-Waterman algorithm to detect and group similar video sub-sequences to find retakes. Our experiments evaluated on the standard benchmark dataset of TRECVID BBC Rushes 2007 show the effectiveness of the proposed method.

Keywords: Retake detection, Video summarization, Rushes video.

1 Introduction

Rushes are raw material (extra video, B-rolls footage) shot during the making of motion picture. During filming section, a large amount of these material may be shot but only a small fraction of them actually become a part in the final product. The reason is that the director always asks a typical scene to be shot for several times if he decides that an additional take is required (e.g. Because the actor gets his lines wrong or an unexpected object suddenly appears in the scene...). Each time of shooting forms one retake of that scene. Retakes of the same scene are often different in duration because the director may stop the recording suddenly if the performance is not going well. Even when the two takes have the same length, they are still slightly different because actions performed in each take are similar, but not identical. An example of retakes in rushes video is shown in Fig 1.

In this research, we deal with the problem of detecting and grouping multiple takes of the same scene in rushes video. This has always been one of the most

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 608–615, 2013.

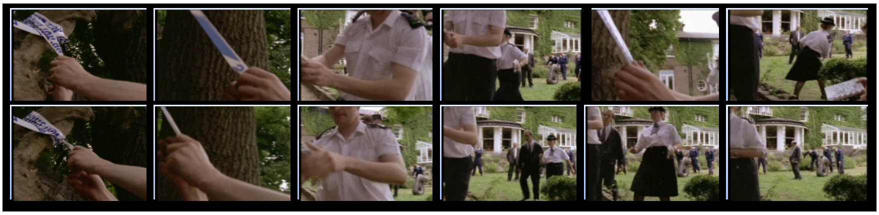
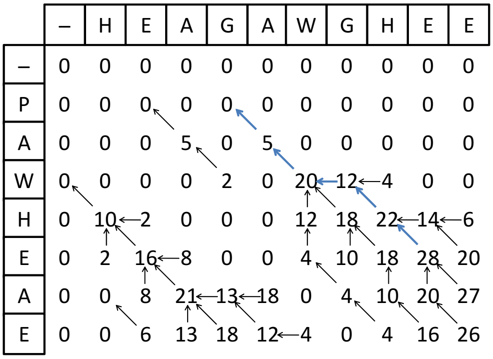


Fig. 1. Two takes of one scene in from video MRS145905

important tasks in the making of a TV product. Retake detection allows editors organizing and structuring rushes video into takes and scenes so that they can identify which fragments of the video should be used for creating the final product. In addition, the video should be indexed so that it is well documented and thus reusable. In order to get a good view, a scene is typically recorded from multiple camera angles. Each camera angle forms one recording of that scene. We restrict our research scope in which each recording from each camera angle is considered as a scene, and each time of recording is one retake of that scene.

Conventional retake detection systems [1,2,3,4] share the common approach consisting of two main steps. In the first step, an input video is partitioned into segments and each segment is quantized into a label of concept. After this step, the video is represent as a sequence of labels. In the second step, similar subsequences are grouped using sequence alignment algorithm. The representation step can be done using some kinds of clustering methods. Similar video segments are grouped into the same cluster label. The labels can be interpreted as the level of similarity, which indicate whether the two video segments are illustrate the same concept or not . However, these methods only work well if the number of labels needed to represent those concepts in video is known in advance. In general case, these approaches usually produce unsatisfactory results.



	–	H	E	A	G	A	W	G	H	E	E
–	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	20	12	4	0	0
H	0	10	2	0	0	0	12	18	22	14	6
E	0	2	16	8	0	0	4	10	18	28	20
A	0	0	8	21	13	18	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

Fig. 2. Retake detection by sequence alignment

In this paper, we introduce a different approach in-which we formulate the representation step as a binary classification problem. Instead of quantizing video segments into a sequence of labels, we propose to use a supervised learning method to classify each pair of segments in the video as matched or mismatched. Then we use this information as an input for aligning all similar sub-sequences in the video. Our proposed classification method has the following advantages. First, since this is a binary classification, there is no need to decide the number of labels beforehand. Second, our method is domain independent,i.e. it does not require domain-specific knowledge of trained data. The details of our work are presented as follows: In section 2 we briefly review related work of retake detection based on sequence alignment methods. In section 3, we explain the details of our proposed method and the framework for retake detection. Section 4 describes the experimental result on TRECVID 2007 BBC Rushes dataset. Finally, we summarize our paper in Section 5.

2 Related Work

Sequence alignment algorithms have widely been used to identify regions of similarity (alignments) between two sequences of proteins. These algorithms are based on the technique of dynamic programming to produce the global alignment via the Longest Common Sub Sequence algorithm or local alignment via the Smith-Waterman algorithm. Because a video can be represented as a sequence of protein, sequence alignment can be applied to detect similar sub-sequences of the video.

Bailer et al.[1] proposed the modified version of Longest Common Subsequence (LCSS) model to measure the similarity distance between different parts of the input video. Each part is determined by shot boundary detection. In [2], Cooharojananone also use LCSS model combine with SIFT feature matching to define the similarity between pairs of pre-segmented shots. Chanasis et al.[5] decompose a video in to multiple shots and then perform a global alignment between all pairs of shots. In [6] Liu et al. also used global alignment to align the sequences of two adjacent sub-shots, and determine whether they are matched or partly matched. However, global alignment approaches always assume that take boundaries are provided in advance by some shot boundary detection tools. In reality, the director may ask a scene to be recorded continuously. Consequently, shot boundary cannot be a good option to find all take boundaries in rushes.

Since repetitive takes of the same scene are represented by the sub-sequences that repeatedly appear in different positions in the video sequence, *local alignment* algorithm such as Smith-Waterman[7] can be used to detect them. Instead of looking at the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure. The example of local alignment can be found in the work of Dumont and Merialdo[3] and Hiep et al.[4]. The Smith-Waterman algorithm builds a scoring matrix with cells represent the cost to change a sub-sequence of the first sequence into a sub-sequence of the second one. The scoring matrix for the two sequences is built as

```
**Input**: A = $a_1a_2a_3...a_m$ and B = $b_1b_2b_3..., b_n$
**Output**: A scoring matrix $M(m+1, n+1)$
$M(i,0) = 0$, $M(0,j) = 0$, $0 \leq i \leq m$, $0 \leq j \leq n$;
$M(i,j) = max \begin{pmatrix} 0 \\ M(i-1,j-1) + w(a_i,b_j) \\ M(i-1,j) + w(a_i,-) \\ M(i,j-1) + w(-,b_j) \end{pmatrix}$
```

Algorithm 1. Calculate Smith-Waterman scoring matrix

in Algorithm 1. Where $w(a_i, b_j)$ is the *match/mismatch* score, if $a_i = b_j$ then $w(a_i, b_j) = w_{match}$ otherwise $w(a_i, b_j) = w_{mismatch}$. $M(i, j)$ denotes the similarity score of the two sequences end at a_i, b_j respectively. Back-tracking starts at the highest cells of the scoring matrix M and propagates until the cell with score zero is encountered, to find the optimal local alignment. This algorithm is applied with the two input sequences are the same video sequence $S = s_1s_2s_3s_4...s_n$ with s_i denotes the label of i^{th} segment. This means we are trying to identify the similar sub-sequences within the video sequence itself. Fig 2 illustrates the example of Smith-Waterman algorithm for detecting repetitive in video.

3 Proposed Method

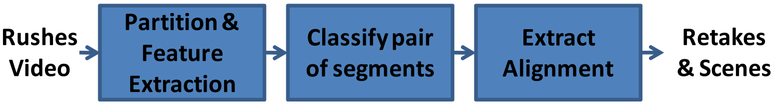


Fig. 3. Process of retake detection in rushes video

This section introduces our frame work for retake detection and our proposed method of binary classification.

3.1 Proposed Framework

The process of our framework is illustrated in Fig 3. Firstly, the input video is partitioned into multiple one-second segments (25 frames). For each segment we choose the middle frame as representative keyframe and extract feature vectors. In the next step, we compute feature between every pairs of segments and use a pre-trained classifier to decide whether the two segments are match or mismatch. The output of this step is a similarity matrix I indicates matching between pairs of segments. Finally, we perform the Smith-Waterman algorithm to detect and group similar sub-sequences to identify retakes and scenes.

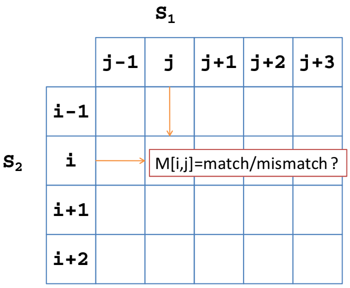


Fig. 4. The similarity matrix between pair of video segments

3.2 Video Partition and Feature Extraction

We use the middle frame of each one-second segment as represent keyframe and form feature vector F based on this keyframe. Each frame is divided into 36 $(= 6 \times 6)$ sub-images. For each sub-image, 24 bin color histogram on HSV color space is extracted. Finally, each segment is represented by an 864-dimensional feature vector.

3.3 Binary Classification Pair of Segments

Typically, the Smith-Waterman algorithm will assign a score for aligning a pair of match/mismatch labels or a gap penalty for aligning a label in one sequence to a gap in the other. The match/mismatch is decided by comparing the quantized labels of the two segments. Each label represents each different concept in rushes video and they are assigned by some quantization methods such as clustering [3,4]. Because the number of scene in video is not given and we can not estimate how much concepts it takes to represent a video, quantization methods always lead to unsatisfactory results.

Actually, to calculate the scoring matrix M, the Smith-Waterman algorithm only wants to know whether the segment at position i is similar (matched) or dissimilar (mismatched) to the segment at position j. This can be formulated as a binary classification problem with match for positive and mismatch for negative. In other to decide whether the two segments s_i and s_j are match or not match, we use a classifier C to compare their correspondent features. The two segments is matched if they are closed together in the vector space, otherwise they are mismatched. Therefore we choose Euclidean distance between two video segments as the feature for each pair. Let $F_i(a_1, a_2, ..., a_n)$ be the feature vector of the first segment and $F_j = (b_1, b_2, ..., b_n)$ be the feature vector of the second one. Let a_k and b_k is the k^{th} component of F_i and F_j respectively. We compute the pair feature using as follow:

$$d(F_i, F_j) = \sqrt{\sum_{k=1}^{n} (a_k - b_k)^2} \tag{1}$$

From the training data, we compute a list pair of segments which is labeled as match or mismatch based on equation (1). Since there is only one input feature, we apply the method of decision stump to train the classifier C. Finally, we create a similarity matrix I where each cell indicates that the segment at position i is matched to the segment at position j.

3.4 Extract Alignment and Form Take and Scene

```
Input: A similarity matrix $I$
Output: A list of aligned sub-sequences
Calculated the Scoring matrix $M$ by algorithm 1;
repeat
    Find position $(i, j)$ where $M(i, j)$ is maximum;
    Trace back to find optimum alignment;
    if $len >= minlen$ then
        Store the alignment;
    end
    Update scoring matrix and find next $M(i, j)$;
until $M(i, j) < threshold$;
```

Algorithm 2. Alignment extraction algorithm

After calculate the similarity matrix I, we perform the Smith-Waterman algorithm to detect all similar sub-sequence in video as in Algorithm 2. The result of this step is a list of candidate takes. Aligned takes, that are extract from the previous step, are grouped to form a scene. We also eliminate all the takes whose length is too short and merge all the overlapping takes. The boundary of a scene is determined by $[min(start_1, ..., start_n), max(end_1, ..., end_n)]$ with $(start_i, end_i)$ are the boundary of take i^{th} take and n is the number of take in that group. The process of forming takes and scene is shown in Fig 5. The first line is the ground truth. The next lines are pairs of alignment between takes. Each pair of alignment is in the same line with the same color. The blue circle illustrates a merger between overlap take candidates. The red circle stands for take candidates whose length is too short. The last line is the final result where open dot rectangles represent the result from merged candidates.

4 Experimental Results

The framework has been evaluated on a subset of TRECVID 2007 BBC Rushes Video Summarization dataset. The dataset is in MPEG-1 format, recorded from about fives BBC dramatic series. Most of the videos have duration of about 30 minutes. Each video is 350 × 288 in resolution and has 25 frames per second. We randomly select 5 videos from the dataset for the experiment.We summarize our dataset in Table 1. We use one video for training and the others for testing.

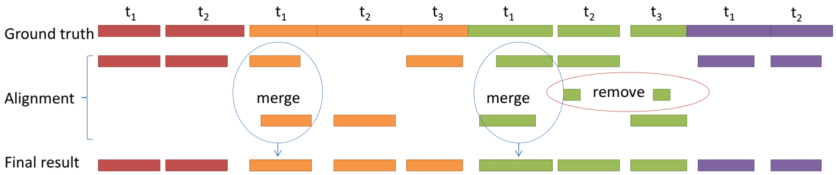


Fig. 5. The process of forming takes and scenes

Table 1. The rushes 2007 database

Video Name	Length(s)	No. Take	No. Scene
MRS025913	1543	28	8
MRS144760	1631	22	6
MRS157475	1557	36	9
MS216210	1453	22	8
MS210470	949	22	10
Average	1402	25	8

For each video we manually annotate the ground truth by identifying the set of scenes and the takes of each scene. We also consider remove the test pattern, color bars and monochrome scenes from the video. Finally, we apply the Rand Index [8] method for evaluation system output against the ground truth.

Our experiment results are shown in Table 2. We compare our result to the common method, which quantize video as a sequence of labels using *k-means*. In the clustering and labeling step, we use different number of clusters $k = 30, 35, 40, 45, ...$ and report the best result as well as the average result of all k. As shown in Table 2, our method is very competitive to the best performance of *k-means* in both scene take detection. However, in some video where scenes are closely similar, e.g. MRS025913, our method does not help improve the performance of scene detection. The reason is that the classifier is not strong enough to reject mismatched pairs from different scenes. Because of that, takes

Table 2. Experiment results in TRECVID 2007

	Avg K-means		Best K-means		Ours	
Video Name	Take RI	Scene RI	Take RI	Scene RI	Take RI	Scene RI
MRS025913	0.69	0.66	0.70	0.67	0.73	0.51
MRS144760	0.89	0.84	0.93	0.88	0.92	0.91
MRS157475	0.81	0.79	0.87	0.86	0.86	0.82
MS216210	0.87	0.84	0.88	0.86	0.86	0.85
MS210470	0.74	0.73	0.72	0.71	0.87	0.85
Average	0.80	0.77	0.82	0.80	0.84	0.79

from different scenes are grouped into the same cluster. Note that because we choose the best *k-means* based on the average Rand Index at each *k*, the best result for some video may less than the overall average (e.g. in video MS210470).

5 Conclusion

In this paper we introduce a new approach for retake detection in rushes video. Most previous detection methods represent input rushes video into a sequence of labels. This might lead to unsatisfactory results because the number of labels needed for a video is always unknown. Instead of represent video segments as a sequence of labels, we build a similarity matrix for matching between pair of segments. We then formulate the matching process between pair of segments as a binary classification problem. Experimental results on BBC Rushes dataset of TRECVID 2007 show effectiveness of the proposed method. In the future, we would like to conduct an evaluation on different type of low-features and their combination for the problem of retake detection.

References

1. Bailer, W., Lee, F., Thallinger, G.: Detecting and clustering multiple takes of one scene. In: Satoh, S., Nack, F., Etoh, M. (eds.) MMM 2008. LNCS, vol. 4903, pp. 80–89. Springer, Heidelberg (2008)
2. Cooharojananone, N., Putpuek, N., Satoh, S., Lursinsap, C.: A novel retake detection using LCS and SIFT algorithm. In: Muneesawang, P., Wu, F., Kumazawa, I., Roeksabutr, A., Liao, M., Tang, X. (eds.) PCM 2009. LNCS, vol. 5879, pp. 777–787. Springer, Heidelberg (2009)
3. Dumont, E., Mérialdo, B.: Rushes video summarization and evaluation. Multimedia Tools Appl. 48(1), 51–68 (2010)
4. Van Hoang, H., Le, D.-D., Satoh, S., Nguyen, Q.H.: Improving retake detection by adding motion feature. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part II. LNCS, vol. 6979, pp. 150–157. Springer, Heidelberg (2011)
5. Chasanis, V., Likas, A., Galatsanos, N.: Video rushes summarization using spectral clustering and sequence alignment. In: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, TVS 2008, pp. 75–79. ACM, New York (2008)
6. Liu, Y., Liu, Y., Ren, T., Chan, K.C.C.: Rushes video summarization using audio-visual information and sequence alignment. In: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, TVS 2008, pp. 114–118. ACM, New York (2008)
7. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology 147(1), 195–197 (1981)
8. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336), 846–850 (1971)

Person Re-identification Using Deformable Part Models

Vu-Hoang Nguyen[1], Kien Nguyen[2], Duy-Dinh Le[1,2],
Duc Anh Duong[1], and Shin'ichi Satoh[2]

[1] Multimedia Communications Laboratory, University of Information Technology,
VNU-HCM, Ho Chi Minh City, Vietnam
{vunh,ducda}@uit.edu.vn,
[2] National Institute of Informatics, Tokyo, Japan
{kienng,ledduy,satoh}@nii.ac.jp

Abstract. Person Re-Identification is the problem of matching people across a network of non-overlapping cameras. One of the challenges is how to match body parts to body parts for comparison between images of two people in the context of different viewpoints as well as deformable human bodies. Existing approaches usually use fixed models to localize body parts or detect human shapes to extract body parts from the shapes. Therefore, it is difficult to change to a new model or structure of body parts. Moreover, those approaches could not deal with multiple human poses simultaneously. We propose a machine learning-based method to extract body parts that is based on Deformable Part Models (DPM). DPM is easy to train and has robust performance. In addition, with DPM, we could use multiple models for multiple human poses concurrently. Experiments on standard dataset ETHZ1 show that the proposed method outperforms state of the art methods.

Keywords: Person Re-Identification, Deformable Part Models, Maximally Stable Colour Regions, Cumulative Matching Characteristic.

1 Introduction

Person Re-Identification (Person Re-Id) is the task of re-recognizing people over a network of non-overlapping cameras. This problem has various real world applications such as: people tracking, surveillance, authentication systems. However, due to the quality of surveillance cameras, we cannot have high resolution images of people, thus, biometric cues such as face or iris recognition do not work in this situation. Furthermore, different cameras can lead to mis-matching between people because they can appear with different appearances from different viewpoints or in any illumination conditions. Beside that, occlusion and background are also challenging characteristics of the problem.

A popular approach for this problem is to extract features for person images and compute distances between them to obtain similarities between images. Based on those similarities, a ranked list is produced reflecting the probabilities

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 616–623, 2013.

of being the target person of gallery images. There are two major approaches for extracting features: (i) extracting various features on the whole human bodies ; or (ii) on human body's parts. Because of the superior performance of the later approach, it has been choosen in many recent works ([1], [2], [3], [4]).

One of the most important problems needs to be solved for this approach is finding methods for localizing body parts, which have to be suitable for such kind of surveillance camera images and extracted features. There are several efforts to divide human bodies into smaller parts but most of them show some limits. For instance, SDALF [1] uses a horizontal axis to divide a person into two parts by trying to maximize the color dissimilarity and minimize the area difference between them. By doing this, there is only one way to decompose a human body. In [2], Gheissari et al. use a model fitting technique to fit a triangulated graph to a human body, then body parts are extracted from the fitted model. This approach could apply only one model at a time, therefore it couldn't model multiple human's poses which are much different from each other (eg. front-view, side-view, etc.), a very typical case in person re-identification datasets.

To deal with these problems, in this paper, we propose a methodology for person re-identification based on machine learning-based deformable part models for human body part localization. The idea is to use Deformable Part Models trained from human datasets (eg. INRIA, PASCAL VOC for human, etc.) to detect human body parts. The model with highest score is selected . HSV Histogram and MSCR [5] are then applied on the detected parts. Final distances between person images are computed to infer a ranked list of gallery images. There are two advantages of our approach. First, because we use learning-based approach for body part localization, we can easily change the number and structure of parts in the training phase. Second, multiple models can be applied simultaneously, thus we can deal with many poses concurrently. The method is evaluated on ETHZ1 dataset, one of the standard datasets for person re-identification [1], [4], [6]. The result shows that our method outperforms SDALF [1], a state of the art method for person re-identification.

The rest of this paper is organized as follows: Section 2 is an overview of related work of this problem, Section 3 presents our method in detail. Experimental results and discussions will be shown in section 4. Finally, section 5 is the conclusion of the paper.

2 Related Work

Works about person re-identification could be divided into two directions: direct approach and metric learning approach. In the direct approach, features are extracted from images and distances between probe image and gallery images are computed based on those features to obtain a ranked list. The problem of this direction is (i) what kind of feature to use, and, (ii) how to extract those features from images. Color, a simple but efficient feature, is widely used in [1],[6],[3]. Meanwhile, shape feature is employed in [3],[7]. Biologically inspired features (BIF) for person re-identification is introduced recently in [6]. Moreover, there are several works using texture such as [8],[9]. To extract features, beside visual features from the whole body, there are several ways which try to extract visual features on each part. For this purpose, a method proposed is to match interest points between 2 images [2]. Another is to divide human body into multiple pre-defined parts such as [1], [2], [3], [4].

Metric learning is a completely different direction. In this direction, machine learning is applied to learn for choosing kinds of feature and how to combine those features to boost the performance. Some works focus on metric learning such as [8],[7],[9]. Our method concentrates on the direct approach, specifically on how to match body parts on two images, thus, we can apply different metric learning methods on it for a better result.

3 Proposed Method

3.1 Framework

This section presents an overview of our framework (Fig. 1). First, human bodies are divided into body parts where the number and structure of parts are predefined. Then, visual features are extracted on those parts and accumulated from part to part, feature by feature to form final features on the whole body. Finally, distances between probe image and gallery images are computed based on distances between the corresponding features. A ranked list of gallery images is produced according to their distances to the probe image.

3.2 Body Part Localization Using Deformable Part Models

In order to detect human body parts, we need a method working with deformable objects, satisfying two criteria: (i) able to work with an arbitrary structure and number of parts, and, (ii) able to deal with diverse poses of human being (Eg. front-view and side-view).

In this work, we use Deformable Part Models (DPM) in [10],[11] for body parts localization. DPM was originally designed for object detection. It is a mixture of multiscale deformable part models. Each model detects objects based on a root filter for the whole object and parts filters for object's parts (see Fig. 2). Those filters are trained from datasets of people using latent SVM (LSVM). By using part filters in DPM, we could obtain locations of detected parts to use for Person Re-Identification. Multiple poses problem could be solved when we use multiple models. The model with highest score, which means the most confident model to the image, is selected. Whenever we need new models for new sets of pre-defined parts, we can train again with new annotations on the dataset.

3.3 Feature Extraction

After locating parts, different types of feature are extracted on each part and then accumulated feature by feature across all parts. Because each feature represents

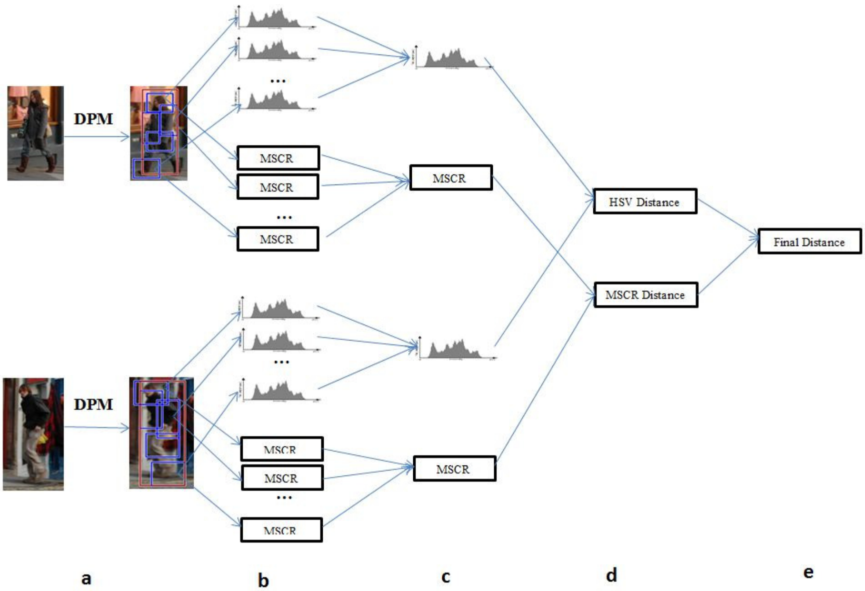


Fig. 1. Computing distance between two images: *(a) Localizing body parts using DPM, (b) Extracting features on detected parts, (c) Accumulating features, (d) Computing partial distances, (e) Computing final distance*

an aspect of a person, combination of multiple features turns out to be efficient in this problem [6]. In [9], Prosser et al. use a combination of 8 color features with 21 texture features ([6]). SDALF [1] fused weighted histogram, Maximally Stable Colour Regions (MSCR), and Recurrent High-Structured Patches (RHSP) and produce a top performance. In [6], the authors combined BIF feature with weighted histogram and MSCR and obtain very good results. In this paper, we also combine 2 kinds of feature which are proved to be among the most efficient kinds of feature.

HSV Histogram. We use color histogram for the 2 reasons: low complexity and high performance. HSV color space is choosen because of its similarity to human's vision system. In our framework, only pixels in foreground regions are counted in the histograms.

Maximally Stable Colour Regions for Recognition and Matching (MSCR). MSCR is a color-based affine covariant region detector [5]. Its outputs are detected regions with their area, centroid, second moment matrix and average color. In this work, only the average colors and centroids are used for the feature extraction.

After obtaining features on each part, HSV histograms of all parts are finally concatenated to form the entire HSV histogram feature for the image. Similarly, average color and centroids of all parts are concatenated to form the entire color and centroids of MSCR of the entire image.

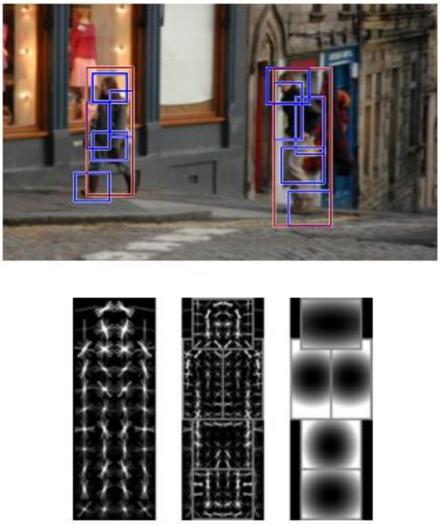


Fig. 2. An example of DPM model [10]

3.4 Computing Distances

In this section, we discuss about how to compute distances between two images. As aforementioned, the final distance between two images must be computed from the HSV Histogram and MSCR features of theirs. Basically, there are two ways of computing final distance: (i) combining all the features of each image to a single feature and compute the distance based on those single features and (ii) computing distances between corresponding features of the two images then the final distance is determined based on those distances. In our method, the later alternative is selected. Specifically, The final distance between two images is a linear combination of the distances between their corresponding features (see Fig. 1).

$$d\left(Img_A, Img_B\right) = \alpha_1 \times d_{HSV}\left(Img_A, Img_B\right) + \alpha_2 \times d_{MSCR}\left(Img_A, Img_B\right) \tag{1}$$

where d_{HSV} refers to distance between HSV Histograms and d_{MSCR} refers to distance between MSCR features. d_{HSV} could be computed by calculating the Bhattacharyya distance between the two histograms. MSCR includes two components: color and centroid. Similar to [1], we only use the y component of the centroid. Accordingly, MSCR distance is a combination of color distance and centroid's y component distance.

4 Experimental Results

In this section, we present our experimental result of the proposed method. To visualize the result, the result is presented by the Cumulative Matching Characteristic

(CMC) curve. A CMC curve [12] represents the probability of the ground truth belonging to top n of the ranked list.

We have evaluated our method on the ETHZ1 dataset [13], one of the standard datasets for person re-identification problem. ETHZ1 consists of 4857 images of 83 people, recorded from a camera attached on a moving charriot. In [7], the dataset were cropped into images of single person by W. R. Schwartz and L.S.Davis. The most challenging characteristics of this dataset are occlusion and illumination changes.

We carry out experiments on single-shot modality in which each person in the gallery set and the probe set is represented by only one image. Because in the ETHZ dataset, there are many images for each person, we randomly choose one image per person to form the gallery set, each of the rest images becomes a probe image. Therefore, there are 83 images coressponding to 83 people in the gallery set and 4774 images in the probe set.

For each experiment, all images must be normalized to a uniform size. With the proposed method, the normalized size is 128 × 64, a suitable size so that DPM works well. For training DPM, INRIA pedestrian dataset with 614 positive samples and 1218 negative samples is choosen as the training dataset. Our DPM model is trained with 1 component and 6 human body parts selected.

The experiment result is shown in Figure 3. We compare our result with SDALF on various sizes: 64 × 32 and 128 × 64. According to the result we can see that SDALF gives higher performance when image sizes are increased and

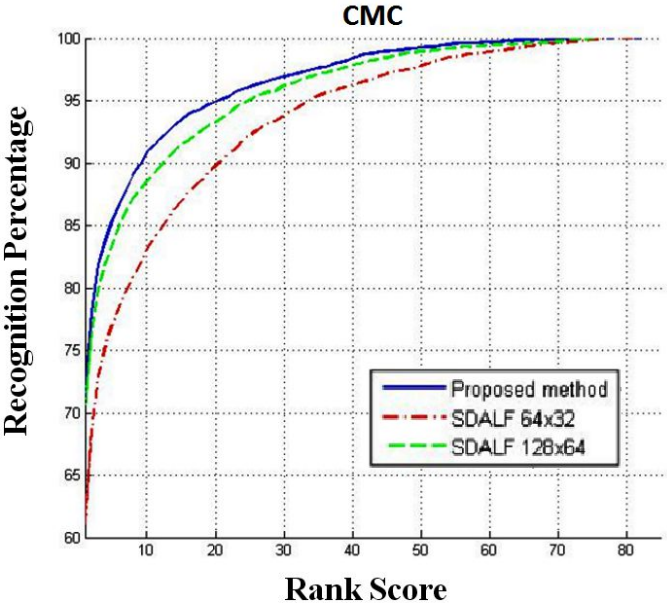


Fig. 3. Experiment result of the propose method, SDALF 128×64, and SDALF 64×32

our method outperforms SDALF in both sizes. The clearest gap of the outperformance is between top 10 and top 20 of the ranked list when the difference fluctuates from 5 to nearly 10 percent (compared to SDALF 64 × 32). The improved performance could be because of the higher accuracy in human parts localization and higher number of parts which can make feature matching more precise. Via this experiment, it turns out that numbers and structures of parts are very important factors affecting the final performance. How to determine the optimal structure is, therefore, a significant problem.

5 Conclusion

We proposed a machine learning-based method for detecting human body parts in the person re-identification problem. This enable us to try with many structures of parts as well as multiple poses concurrently. This approach could be applied on re-identification problem of not only people but also arbitrary deformable objects. The experiment on ETHZ1 shows that, with our new structure of parts, we could obtain a higher performance in person re-identification than SDALF, a method with a fixed technique for detecting human parts. For the future work, adaptively selecting the number of body parts is still a promising direction.

Acknowledgement. This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DHQG-C20132601.

References

1. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010). IEEE Computer Society Press, San Francisco (2010)
2. Gheissari, N., Sebastian, T., Hartley, R.: Person reidentification using spatiotemporal appearance. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1528–1535 (2006)
3. Oreifej, O., Mehran, R., Shah, M.: Human identity recognition in aerial images. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 709–716 (2010)
4. Cheng, D.S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of the British Machine Vision Conference, pp. 68.1–68.11. BMVA Press (2011), http://dx.doi.org/10.5244/C.25.68
5. Forssen, P.E.: Maximally stable colour regions for recognition and matching. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8 (2007)
6. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: Proceedings of the British Machine Vision Conference, pp. 57.1–57.11. BMVA Press (2012)

7. Schwartz, W., Davis, L.: Learning Discriminative Appearance-Based Models Using Partial Least Squares. In: Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing (2009)
8. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
9. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference, pp. 21.1–21.11. BMVA Press (2010), doi:10.5244/C.24.21
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
11. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.: Discriminatively trained deformable part models, release 5, `http://people.cs.uchicago.edu/~rbg/latent-release5/`
12. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS (2007)
13. Ess, A., Leibe, B., Schindler, K., van Gool, L.: A mobile vision system for robust multi-person tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008). IEEE Press (2008)

Convolutional Neural Networks Learn Compact Local Image Descriptors

Christian Osendorfer, Justin Bayer, Sebastian Urban,
and Patrick van der Smagt

Technische Universität München, Fakultät für Informatik, Lehrstuhl für Robotik
und Echtzeitsysteme, Boltzmannstraße 3, 85748 München
osendorf@in.tum.de, bayer.justin@googlemail.com, surban@tum.de, smagt@tum.de

Abstract. We investigate if a deep Convolutional Neural Network can learn representations of local image patches that are usable in the important task of keypoint matching. We examine several possible loss functions for this correspondance task and show emprically that a newly suggested loss formulation allows a Convolutional Neural Network to find compact local image descriptors that perform comparably to state-of-the-art approaches.

Keywords: Convolutional Neural Networks, Non-linear Dimensionality Reduction, Local Image Descriptor Learning.

1 Introduction

Local image descriptors are an important component of many Computer Vision algorithms. They are central to a wide range of Computer Vision tasks like tracking, stereo vision, panoramic stitching, structure from motion or object recognition. Given these widely differing types of use cases, a local image descriptor should be invariant to image, appearance, viewpoint and lightning variations of a local image patch.

Over the last decade many different descriptors have been developed. Several of these are hand designed, with SIFT [1] being the most popular example. In recent years these engineered descriptors where accompanied by approaches that are based on discriminant learning techniques [2–5]. The general motivation behind these methods is that by exploiting statistical properties of image patches through learning, the resulting descriptor is more robust to the previously mentioned variations an image patch can be exposed to.

Tracing back our work to [6], this paper tries to extend the recent success story of Convolutional Neural Networks [7–9] to learning compact local image descriptors. In a series of experiments, we investigate various aspects (different cost functions, different non-linearities, depth) of models based on Convolutional Neural Networks. It turns out that with the correct cost function, Convolutional Neural Networks find compact image descriptors that perform competitively or even better than state-of-the-art algorithms on a challenging benchmark for keypoint matching [3].

M. Lee et al. (Eds.): ICONIP 2013, Part III, LNCS 8228, pp. 624–630, 2013.

Related Work. Similar to our work, [3–5] rely on supervised learning approaches to find compact local image descriptors. These methods suggest different pooling and selection strategies of gradient-based features to learn discriminant descriptors, utilizing boosting [4] or sparse convex optimization [5].

Most similar to our work is [10]. Like us [10] uses a Convolutional Neural Network to learn an encoding for an image patch. However, [10] investigated the applicability of their learnt descriptors only for planar transformations and only compared their performance to SIFT. As it turns out the objective function to train the whole model used in [10] would not be competitive to state-of-the-art approaches on the challenging dataset used in our paper. And finally, [10] relies on gradient based input features on various scales while our algorithm works directly on pixel intensities.

2 General Learning Architecture

A good description of a local image patch is characterized by the fact that *corresponding* image patches are represented by descriptors that are close-by under some metric. *Correspondence* is thereby defined by the various kinds of invariances listed in the first paragraph of section 1. Clearly, the goal of any learning algorithm in this domain is then to find representations *together* with the accompanying metric that performs well on labeled image pairs (corresponding vs. non-corresponding pairs).

DrLim [6] is a framework for energy based models that learn representations using only such correspondence relationships. We utilize DrLim in order to learn low-dimensional mappings for low-level image patches.

The main idea behind DrLim is to map similar (i.e. corresponding) image patches to nearby points on the output manifold and dissimilar image patches to distant points. DrLim is defined over pairs of image patches, x_1 and x_2. The i-th pair (x_1^i, x_2^i) is associated with a label y^i, with $y^i = 1$ if x_1^i and x_2^i are deemed similar and $y^i = 0$ otherwise. We denote by $d(x_1, x_2; \theta)$ the parameterized distance function between the representations of x_1 and x_2 that we want to learn. Based on $d(x_1, x_2; \theta)$ we define DrLim's loss function $\ell(\theta)$:

$$\ell(\theta) = \sum_i y^i \ell_{\text{pll}}(d(x_1^i, x_2^i; \theta)) + (1 - y^i)\ell_{\text{psh}}(d(x_1^i, x_2^i; \theta)) \tag{1}$$

We denote with $\ell_{\text{pll}}(\cdot)$ the partial loss function for similar pairs (it *pulls* similar pairs together) and with $\ell_{\text{psh}}(\cdot)$ the partial loss function for dissimilar pairs (it *pushes* dissimilar pairs apart). Several possible choices for $\ell_{\text{pll}}(\cdot)$ and ℓ_{psh} (denoted by C_i) are investigated in this text:

- C_1 — the original paper for DrLim [6] defined $\ell_{\text{pll}}(\cdot)$ and ℓ_{psh} as follows:

$$\ell_{\text{pll}}(d(x_1, x_2; \theta)) = c_{\text{pll}} d(x_1, x_2; \theta)^2 \tag{2}$$

$$\ell_{\text{psh}}(d(x_1, x_2; \theta)) = c_{\text{psh}}[\max(0, m_{\text{psh}} - d(x_1, x_2; \theta))]^2 \tag{3}$$

m_{psh} is a push *margin*: Dissimilar pairs are not pushed farther apart if they already are at a distance greater than m_{psh}. c_{pll} and c_{psh} are scaling factors, both set to $\frac{1}{2}$ in [6].

- C_2 — [10] uses the definitions from [11]:

$$\ell_{\text{pll}}(d(x_1, x_2; \theta)) = \frac{2}{Q} d(x_1, x_2; \theta)^2 \tag{4}$$

$$\ell_{\text{psh}}(d(x_1, x_2; \theta)) = 2Q \exp(-\frac{2.77}{Q} d(x_1, x_2; \theta)) \tag{5}$$

The constant Q is set to the upper bound of $d(x_1, x_2; \theta)$.

- C_3 — the exponential loss from [4]:

$$\ell_{\text{pll}}(d(x_1, x_2; \theta)) = \exp(y' d(x_1, x_2; \theta)) \tag{6}$$

$$\ell_{\text{psh}}(d(x_1, x_2; \theta)) = \exp(y' d(x_1, x_2; \theta)) \tag{7}$$

where $y' = 2y - 1$ and y indicates whether a given x_1 and x_2 are a corresponding pair or not, i.e. $y' \in \{-1, 1\}$

- C_4 — in this paper we investigate a combination of a hinge-like loss function for ℓ_{pll} with ℓ_{psh} set as in [6]:

$$\ell_{\text{pll}}(d(x_1, x_2; \theta)) = c_{\text{pll}}[\max(0, d(x_1, x_2; \theta) - m_{\text{pll}})] \tag{8}$$

$$\ell_{\text{psh}}(d(x_1, x_2; \theta)) = c_{\text{psh}}[\max(0, m_{\text{psh}} - d(x_1, x_2; \theta))]^2 \tag{9}$$

m_{pll} is a pull *margin*: Similar pairs are pulled together only if they are at a distance above m_{pll}.

For a complete definition of $\ell(\theta)$ we still need $d(x_1, x_2; \theta)$: for C_1, C_3 and C_4 it is defined as the Euclidean distance between the learned representations of x_1 and x_2:

$$d(x_1, x_2; \theta) = \|f(x_1; \theta) - f(x_2; \theta)\|_2 \tag{10}$$

For C_2 it is defined as

$$d(x_1, x_2; \theta) = \|f(x_1; \theta) - f(x_2; \theta)\|_1 \tag{11}$$

In both cases $f(\cdot)$ denotes the mapping from the (high-dimensional) input space to the low-dimensional representation space. In this paper, $f(\cdot)$ is a Convolutional Neural Network.

2.1 Convolutional Neural Networks

A Convolutional Neural Network [7] is a special kind of neural network for working with images. It is composed of multiple layers, where the output of every layer is a set of two dimensional arrays called feature maps. A feature map is produced by convolving the respective input with a filter, followed by a non-linear function and a pooling layer. Within the DrLim framework the same network is applied to two different inputs in order to compute the loss for this input pair (see equation 10). Therefore, the architecture is sometimes called a siamese network [12, 13]. In this work we investigate two aspects of a configuration of a Convolutional Neural Network:

- non-linearities: we compare the standard $\tanh(\cdot)$ and the currently often used rectifying linear unit [9].
- depth: we compare models with three and four layers.

3 Experiments

We use the dataset from [3] for evaluating various instances of Convolutional Neural Networks. In contrast to previous approaches actual 3D correspondences, obtained via a stereo depth map, are used for generating this dataset. This allows learning descriptors that are optimized for the non-planar transformations and illumination changes that result from viewing a truly 3D scene. The dataset is based on more than 1.5 million image patches (64×64 pixels) of three different scenes: the Statue of Liberty (about 450,000 patches), Notre Dame (about 450,000 patches) and Yosemites Half Dome (about 650,000 patches). We denote these scenes with LY, ND and HD respectively. There are 250000 corresponding image patch pairs and 250000 non-corresponding image patch pairs available for every scene. We train on one scene and evaluate the learned embedding function on the other two scenes. Evaluation is done on the same test sets (50000 matching and non-matching pairs) used also by other approaches.

We achieve the best results on this benchmark with a Convolutional Neural Network paired with the loss function C_4. The network has 4 convolutional layers[1] and uses the tanh non-linearity. Moreover, Table 1 shows that this Convolutional Neural Network (the entry denoted *CNN1*) performs comparably to other state-of-the-art approaches in terms of the 95% error rate which is the percent of incorrect matches when 95% of the true matches are found: After computing the respective distances for all pairs in a test set, a threshold is determined such that 95% of all matching pairs have a distance below this threshold. Non-matching pairs with a distance below this threshold are considered incorrect matches. Figure 1 shows the ROC curves of *CNN1* for the three different training settings.

In order to avoid unnecessary clutter, we describe only qualitatively the results of comparing different settings for loss functions, non-linearities and depth:

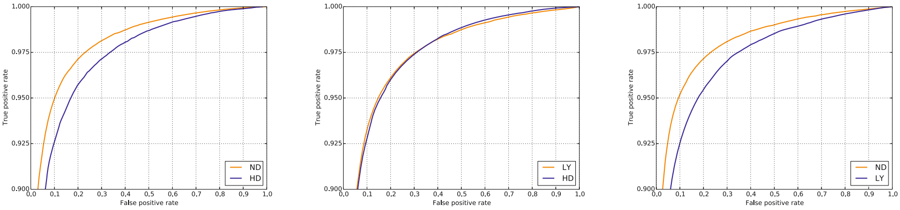
[1] 20 feature maps with kernel size 5×5 followed by a $(2, 2)$ max pooling; a second convolutional layer, again with 20 feature maps and kernel size 5×5 and $(2, 2)$ max pooling; a third convolutional layer, again with 20 feature maps and kernel size 4×4 and $(2, 2)$ max pooling; and a fourth convolutional layer with 64 feature maps and kernel size 5×5.

- Loss functions: C_4 performed at least by $2\% - 3\%$ better than C_1, C_2 or C_3. The idea of having a *pull* margin m_{pll} is crucial for the good performance of C_4. Without it, a noticeable performance drop happens. Interestingly, the results from the original DrLim formulation (C_1) can be improved by utilizing a pull margin, too.
- Non-linearities: Contrary to recent reports [9] on good performance due to linear rectifying units, the networks with a tanh non-linearity performed at least by 5% better than those a the linear rectifying unit.
- Depth: We also tested a Convolutional Neural Network with 3 layers (the total number of parameters was similar to the network with 4 layers). The 4 layer network outperformed this network by approximately $1\% - 1.5\%$.

Table 1. Error rates, i.e. the percent of incorrect matches when 95% of the true matches are found. Every subtable, indicated by an entry in the *Method* column, denotes a descriptor algorithm. The line below every method denotes the size of the descriptor (e.g. 32d denotes a 32 dimensional descriptor). The 128 dimensional SIFT descriptor [1] does not require learning (denoted by – in the column *Training set*). The numbers in the columns labeled LY, ND and HD are the error rates of a method on the respective test set for this scene. [3, 5] do not have results when trained on the LY scene (indicated by ×). L-BGM is presented in [4]. *CNN2* is trained on *two* out of the three datasets, see section 3.1

		Test set		
Method	**Training set**	LY	ND	HD
SIFT	–	31.7	22.8	25.6
L-BGM (64d)	LY	–	14.1	19.6
	ND	18.0	–	15.8
	HD	21.0	13.7	–
[3] (29d)	LY	–	×	×
	ND	16.8	–	13.5
	HD	18.2	11.9	–
[5] (29d)	LY	–	×	×
	ND	14.5	–	12.5
	HD	17.4	9.6	–
CNN1 (32d)	LY	–	10.1	17.6
	ND	14.6	–	15.3
	HD	17.6	9.5	–
CNN2 (32d)	LY/ND	–	–	12.3
	LY/HD	–	7.3	–
	ND/HD	13.3	–	–

Every image patch is preprocessed by subtracting its mean and dividing by its standard deviation. All models are trained with standard gradient descent. Training stops when a local minimum of the DrLim objective is reached. We never faced the problem of overfitting (probably because the number of parameters is very small compared to the size of the training set), and thus did not use a validation set. Instead we observed that using a validation set had a negative effect on our final results – the data in the validation set is more useful for actual training. Finally, the hyperparameters for C_4, namely c_{pll}, c_{psh}, m_{pll} and m_{psh} are 0.5, 3, 1.5, and 5 respectively. Notably, these hyperparameters are *not* scene dependent.



(a) Training set: LY

(b) Training set: ND

(c) Training set: HD

Fig. 1. True Positive Rates and False Positive Rates for *CNN1*. A plot is denoted by its training set and shows the ROC curves on the two remaining test sets. Best viewed in color.

3.1 Data Augmentation

Convolutional Neural Networks benefit from abundant data. A successful method to artificially enlarge the available amount of data is to generate new input data by applying different kinds of transformations to the original dataset [8, 9]. Yet, we did not manage to improve the error rates that we achieve on the original dataset with this approach. However, utilizing data from two scenes improves error rates noticeably: we train on two scenes and evaluate on the remaining one. Following this approach, we are able to improve our error rates by at least 2% (see Table 1, last entry, *CNN2*).

4 Conclusion and Future Work

In this short paper we showed empirically that a standard Convolutional Neural Network, equipped with a suitable loss function, can find compact representations for local image patches: on a challenging dataset for keypoint matching we were able to perform at least as well as state-of-the-art approaches.

The appeal of a simple parametric model like a Convolutional Neural Network is that it does not require any complex parameter tuning or pipeline optimizations and that it can be integrated into larger systems that can then be trained in an end-to-end fashion [14]. To be more concrete, the 32 dimensional descriptor proposed in this paper can be used to define a dense representation of an arbitrary image. This dense representation is then fed into another Convolutional Neural Network for e.g. image segmentation [15], which can tune the low-level representations for the specific task at hand through straightforward backpropagation.

Acknowledgments. Sebastian Urban was supported by German Research Foundation (DFG) SPP 1527 Autonomes Lernen.

References

1. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
2. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: Proc. CVPR (2008)
3. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. Pattern Analysis and Machine Intelligence 33(1), 43–57 (2010)
4. Trzcinski, T., Christoudias, M., Lepetit, V., Fua, P.: Learning image descriptors with the boosting-trick. In: Proc. NIPS (2012)
5. Simonyan, K., Vedaldi, A., Zisserman, A.: Descriptor learning using convex optimisation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 243–256. Springer, Heidelberg (2012)
6. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proc. CVPR (2006)
7. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multistage architecture for object recognition? In: Proc. ICCV (2009)
8. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Proc. CVPR (2012)
9. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proc. NIPS (2012)
10. Jahrer, M., Grabner, M., Bischof, H.: Learned local descriptors for recognition and matching. In: Computer Vision Winter Workshop (2008)
11. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proc. CVPR (2005)
12. Becker, S., Hinton, G.E.: Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355(6356), 161–163 (1992)
13. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using siamese time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence 7(4), 669–688 (1993)
14. Hadsell, R.: Learning long-range vision for an offroad robot. PhD thesis, New York University (2008)
15. Schulz, H., Behnke, S.: Learning object-class segmentation with convolutional neural networks. In: Proc. ESANN (2012)

Author Index