Proceedings of

2018 the 2nd International Conference on Video and Image Processing (ICVIP 2018)

December 29-31, 2018

Hong Kong

The Association for Computing Machinery
2 Penn Plaza, Suite 701
New York New York 10121-0701

ACM ISBN: 978-1-4503-6613-7

2018 the 2nd International Conference on Video and Image Processing (ICVIP 2018)

Table of Contents

Preface……...……... …vii

Conference Committees…….....….....viii

Chapter 1: Pattern Recognition

Additive Margin Softmax with Center Loss for Face Recognition 1

Mingchao Jiang, Zhenguo Yang, Wenyin Liu and Xiaochun Liu

A New Method for Stroke Order Recognition of Handwritten Chinese Characters 7

Junjian Tang and Jun Guo

Kick Recognition System Using Dual RGB-D Cameras 12

Sungjin Hong, Myunggyu Kim and Yejin Kim

LoID-EEC: Localizing and Identifying Early Esophageal Cancer Based on Deep Learning in Screening Chromoendoscopy 17

Xiaoxiao Du, Ya Li, Jianning Yao, Bing Chen, Jiayou Song and Xiaonan Yang

Automatic Recognition for Arbitrarily Tilted License Plate 23

Nanxue Lu, Wei Yang, Ajin Meng, Zhenbo Xu, Huan Huang and Liusheng Huang

Real-Time Face Attendance Marking System in Non-cooperative Environments 29

Kai Jin, Xuemei Xie, Fangyu Wang, Xu Gao and Guangming Shi

Chapter 2: Target Detection

Top View Person Detection and Counting for Low Compute Embedded Platforms 35

Prashant Maheshwari, Doney Alex, Sumandeep Banerjee, Saurav Behera and Subrat Panda

Image Edge Detection using Fractional Order Differential Calculus 44

S. Gupta, A. Bhardwaj, R. Sharma, P. Varshney and S. Srivastava

Real-time Anomaly Detection with HMOF Feature 49

Huihui Zhu, Bin Liu, Yan Lu, Weihai Li and Nenghai Yu

Cloud Detection for DSCOVR EPIC Data 55

Qing Guo, Yanhong Zhao and An Li

A FPN-Based Framework for Vehicle Detection in Aerial Images 60

Yinjuan Gu, Bin Wang and Bin Xu

DSBI: Double-Sided Braille Image Dataset and Algorithm Evaluation for Braille Dots Detection 65

Renqiang Li, Hong Liu, Xiangdong Wang and Yueliang Qian

A Natural Scene Edge Detection Algorithm Based on Image Fusion 70

Weichao Ding, Zhile Yang and Liangbing Feng

Optimisation of Feature Space for People Detection from TopView on Light Embedded Platform 75

Doney Alex, Prashant Maheshwari, Sumandeep Banerjee and Subrat Panda

Chapter 3: Image 3D Reconstruction

3D Face Reconstruction from Low-Resolution Images with Convolutional Neural Networks 83

Rouven Winkler, Chengchao Qu, Sascha Voth and Jürgen Beyerer

Planetary Marching Cubes: A Marching Cubes Algorithm for Spherical Space 89

Zackary P. T. Sin and Peter H. F. Ng

Similarity Analysis of 3D Models Based on Convolutional Neural Networks with Threshold 95

Shengwei Qin, Zhong Li and Zihao Chen

A Simple Image Acquisition System and Its Calibration for Image-Based 3D Reconstruction 103

Xiaoming Wang, Limin Shi and Huarong Xu

Chapter 4: Image Analysis

Omnidirectional Saliency Map Generation by Yin-Yang Grid Method 108

Daiki Okazaki, An-shui Yu and Kenji Hara

An Efficient Non-convex Mixture Method for Low-rank Tensor Completion 112

Shi Chengfei, Huang Zhengdong, Wan Li and Xiong Tifan

Clustering color segmentation in multi-color space 118

Ting Tu, Zhifeng Zhou and Peng Xiao

A Hybrid DCT-CLAHE Approach for Brightness Enhancement of Uneven-illumination Underwater Images 123

Meng Ge, Qingqing Hong and Lifeng Zhang

Exemplar-Based Image Inpainting using Automatic Patch Optimization 128

Xuehui Bi, Huaming Liu, Guanming Lu, Jian Wei and Yan Chao

Analysis of Chromatic Characteristics, in Satellite Images for the Classification of Vegetation Covers and Deforested Areas 134

Wilver Auccahuasi, Madelaine Bernardo, Elizabeth Oré Núñez, Fernando Sernaque*, Percy Castro and Luis Raymundo*

Anchor Graph Hashing with Feature Learning 140

Weiguang Li and Xiaogang Peng

Robust Weighted Keypoint Matching Algorithm for Image Retrieval 145

Da-Mi Jeong, Ji-Hae Kim, Young-Woon Lee and Byung-Gyu Kim

Image Retrieval Method with Fisher GLCM and Graph Model 150

Honghu Hua, Jianghua Cheng and Tong Liu

Chapter 5: Video Processing Technology and Method

A Long-Short Term Memory Neural Network Based Rate Control Method for Video Coding 155

Zheng-Teng Zhang, Jucai Lin, Ruidong Fang, Juan Lu and Yao Chen

Improvement on Demosaicking in Plenoptic Cameras by Use of Masking Information 161

Hyunji Cho, Joungeun Bae, Hyeonah Jung, Eunju Oh and Hoon Yoo

MLN: Moment localization Network and Samples Selection for Moment Retrieval 165

Bo Huang, Ya Zhang and Kai Yu

A New Multi-Camera Dataset with Surveillance, Mobile and Stereo Cameras for Tracking, Situation Analysis and Crime Scene Investigation Applications 171

Thomas Pollok

A Fast Block Partitioning Algorithm Based on SVM for HEVC Intra Coding 176

Jun Yin, Xu Yang, Jucai Lin, Yao Chen and Ruidong Fang

Chapter 6: Image Processing Technology and Application

High Signal to Noise Ratio Weld Pool Imaging Device Research in CMT+P 182

Pan Yang, Zhuang Zhao, Jing Han, Yi Zhang and Lianfa Bai

Design a Real-Time Eye Tracker 187

Sara Bilal, Mohamad Hassrol bin Mat Hussin and Rasheed Nassr

Using Bezier Curves to Refine Road Vector Data through Satellite Images 192

Maneesha Perera, Damith Karunaratne and Enosha Hettiarachchi

Single Depth Map Super-resolution with Local Self-similarity 198

Xiaochuan Wang, Kai Wang and Xiaohui Liang

Object Recognition Using Deep Neural Network with Distinctive Features 203

Hyun Chul Song, Farhan Akram and Kwang Nam Choi

Intelligent Information Systems and Image Processing: A Novel Pan-Sharpening Technique Based on Multiscale Decomposition 208

Ahmad Al Smadi and Ahed Abugabah

Dual-band Welding Speed Monitoring Method Based on Deep Learning 213

Jionghang Shen, Zhuang Zhao, Jing Han, Yi Zhang and Lianfa Bai

Template Attentional Siamese Network for Object Tracking 218

Junyan Gao, Zhenguo Yang and Wenyin Liu

Enhanced Satellite Imaging Algorithm Classifier using Convolutional Neural modified Support Vector Machine (CNMSVM) 222

Edgar Bryan B. Nicart, Tony Y.T. Chan and Ruji P. Medina

Brain Tumor Segmentation on MR Images Using Anisotropic Deeply Supervised Convolutional Neural Network 226

Md Minhazul Islam, Zhijie Wang, Muhammad Ather Iqbal and Guangxiao Song

Automatic Prediction of the Conversion of Clinically Isolated Syndrome to Multiple Sclerosis Using Deep Learning 231

H. M. Rehan Afzal, Suhuai Luo, Saadallah Ramadan, Jeannette Lechner-Scott and Jiaming Li

Similarity Detection of Color Image based on Main Color Table 236

Wenjia Ding, Yi Xie and Yulin Wang

Quality Monitoring in Wire-Arc Additive Manufacturing Based on Spectrum 240

Yiting Guo, Zhuang zhao, Jing Han and Lianfa Bai

Preface

This issue of Proceedings gathers the papers presented at 2018 the 2nd International Conference on Video and Image Processing held on December 29-31, 2018 in Hong Kong. ICVIP was initiated in 2017, which is an international conference covering research and development in the field of video and image processing and allow the participation from all over the world.

80 papers were submitted on the conference, and ICVIP finally accepted 45 papers after a double blinded peer review process by international reviewers and technical program committee members. Divided into 6 chapters, the papers provide a wide spectrum of researches on wide range of video and image processing.The chapters are devoted to the framework of pattern recognition, target detection, image 3d reconstruction, image analysis, video processing technology and method as well as image processing technology and application. Specific research results by conference participants were presented and examined in the light of the frameworks outlined above, which is of interest to academics, researchers and professionals in this field.

Four keynote speeches were also presented from Prof. David Zhang, IEEE and IAPR Fellow, from Chinese University of Hong Kong, Shenzhen, China; Prof. Guo Song, from The Hong Kong Polytechnic University, Hong Kong; Prof Mao Kezhi, from Nanyang Technological University, Singapore; Prof. Chin-Chen Chang, IEEE and IET Fellow, from Feng Chia University, Taiwan. All the talks were very impressive for the high level of professionalism, and in many cases original ideas and activities have been accomplished or proposed.

The credit for the success of the conference is to be shared with many colleagues. First and foremost, the advisory chair, conference chairs, program chairs, technical program committee members gave precious inputs and were always side by side with the organizers. We are also indebted to session chairs, international reviewers, conference secretariat who dedicated to make the conference run smoothly and properly, and ensure the proceedings quality. Last but not the least, we should express our thanks to all delegates, who showing the high level of international interest in the subject. It is exactly your participation that make the conference to its success.

The Proceedings provide the permanent record of what were presented. It indicated the state of development at the time of writing of all aspects of this important topic and will be invaluable to all researchers in the field for that reason. We truly believe the participants will find the discussion fruitful, and we hope you enjoy and find your engagement with their ideas valuable in sustaining your own professional development in the field of video and image processing.

Yours Sincerely!

Program Chair
Assoc. Prof. Mao Kezhi
Nanyang Technological University, Singapore

Committees

Conference Chairs
Prof. Chin-Chen Chang, IEEE and IET Fellow, Feng Chia University, Taiwan
Prof. David Zhang, IEEE and IAPR Fellow, Hong Kong Polytechnic University, Hong Kong

Program Chairs
Assoc. Prof. Mao Kezhi, Nanyang Technological University, Singapore
Assoc. Prof. Xuefeng Liang, Kyoto University, Japan

Technical Committee
Prof. Cheng Tin Gan, Florida International University, USA
Assis. Prof. Mahipal Jetta, Mahindra Ecole Centrale, India
Dr. Zhilei Liu, Tianjin University, China
Lecturer Cuicui Zhang, Tianjin University, China
Assit. Prof. Ramesh H, National Institute of Technology Karnataka, India
Prof. Xuemei Xie, Xidian University, China
Prof. Yongmei Zhang, North China University of Technology, China
Prof. Jiancheng Lv, Sichuan University, China
Prof. Xueming Li, Beijing University of Posts and Telecommunications, China
Assoc. Prof. Agus Harjoko, Universitas Gadjah Mada, Indonesia
Prof. Joel Ilao, Macario Cordel Ii, De La Salle University, Philippines
Senior Lecturer Hamid A. Jalab, University of Malaya, Malaysia
Dr. P.Raviraj, Kalaignar Karunanidhi Institute of Technology, India
Dr. Zhang Yu, East-West University, USA
Prof. Surekha Kamath, Manipal Institute of Technology Manipal, India
Prof. Mohammad Shorif Uddin, Jahangirnagar University, Bangladesh
Asst. Prof. Desmond Kho Teck Kiang, Multimedia University, Malaysia
Prof. Rajneesh Sharma, NSIT, India
Dr. Sara Bilal, Whitireia New Zealand, New Zealand
Asst. Prof. Rasheed Nassr, UniKL Malaysian Institute of Information Technology, Malaysia
Dr. Damith Karunaratne, University of Colombo School of Computing, Sri Lanka
Prof. Byung-Gyu Kim, Sookmyung Women's University, South Korea
Asst. Prof. Yejin Kim, Hongik University, Korea
Dr. Ahed Abugabah, Zayed University, United Arab Emirates
Dr. Peter H. F. Ng, The Hong Kong Polytechnic University, Hong Kong
Lecturer Shengwei Qin, Guangzhou University, China
Assoc. Prof. Hong Liu, Chinese Academy of Sciences, China
Assoc. Prof. Liangbing Feng, Chinese Academy of Sciences, China
Assoc. Prof. Xiaonan Yang, Zhengzhou University, China
Prof. Xiaofei Zhu, Chongqing University of Technology, China
Asst. Prof. Edgar Bryan Nicart, Technological Institute of the Philippines, Philippines

Chapter 1: Pattern Recognition

Additive Margin Softmax with Center Loss for Face Recognition

Mingchao Jiang
Guangdong University of Technology, Guangzhou, China
+86 13246876815
13246876815@163.com

Zhenguo Yang*
Guangdong University of Technology, Guangzhou, China
+86 18320145447
yzgcityu@gmail.com

Wenyin Liu*
Guangdong University of Technology, Guangzhou, China
+86 13910586417
liuwy@gdut.edu.cn

Xiaochun Liu
Guangdong University of Technology, Guangzhou, China
+86 18975878383
lxc1448@163.com

ABSTRACT
In this paper, we propose a Convolutional Neural Network(CNN) model by combining Additive Margin Softmax with Center Loss for face recognition, which is denoted as AMS-CL. AMS-CL adopts the CNN network structure of ResNet50, which is pretrained on the FaceScrub dataset. In particular, AMS-CL exploits the Additive Margin Softmax (i.e., AM-softmax) to increase inter-class distance, avoiding overconfidence on face classifications. Furthermore, a regularization term for center loss is introduced by AMS-CL to decrease the intra-class distance simultaneously. Therefore, AMS-CL keeps beneficial inter-class and intra-class relations to classification tasks jointly. Alternatively, we propose to use classifiers on the hidden data representation achieved by AMS-CL to make predictions. Extensive experiments conducted on the LFW dataset show the effectiveness of the proposed AMS-CL model.

CCS Concepts
• **Computing methodologies→Object recognition**

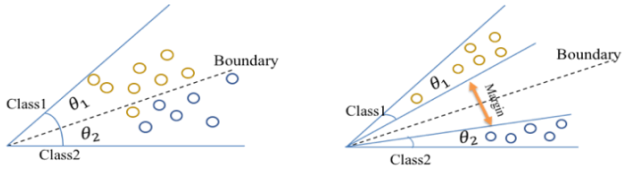
Keywords
Face recognition, AMS-CL, Resnet, Class distance.

1. INTRODUCTION
Face recognition has been widely applied in security verification, entrance and exit control, e-commerce, and many other fields. Because of its friendliness, non-invasion, and easy accessibility, face recognition has attracted a lot of research res with high discriminative power and robustness, and how to attention. At the same time, how to extract facial featu build efficient and reliable classifiers to improve the accuracy of face recognition still are challenging issues.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301511

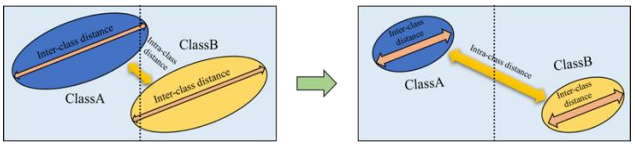
Most of the works on face recognition focus on improving the performance by introducing specific loss functions. Quite a few loss functions have been proposed in recent years, such as softmax for classification, Triplet loss [1], etc. In particular, softmax combined with CNN is widely used in image classification and recognition. However, it usually suffers from overconfidence, which can lead to misclassifications. To address this problem, A-softmax [2] and L-softmax [3] have been proposed. However, these models are difficult to converge in the training stage in practice. As a result, Wang et al. [4] proposed an Additive Margin softmax (known as AM-softmax) for face recognition. AM-softmax can be used by replacing the original softmax layer of the CNN network directly, which normalizes the weight and output of the layer. Figure 1 shows a simple example of softmax and AM-softmax on classifying the data samples into two categories. AM-softmax aims to increase the inter-class distance by maximizing the margin between the categories. The constraint on class boundary of AM-softmax is stronger than the conventional softmax, which can benefit to classification.



(a)Softmax (b)Am-softmax

Figure 1. An intuitive example for softmax and AM-softmax

However, AM-softmax does not take into account the spatial distribution of classes as shown in Figure 2-a. More specifically, the intra-class distance using softmax can be much larger than the inter-class distance, resulting to collisions between classes. Ideally, it is supposed that inter-class distance is much larger than intra-class distance as shown in Figure 2-b. It is significant and expected to reduce the intra-class distance in order to improve the independence of the spatial distribution.



(a)AM-Softmax (b) AM-Softmax with center loss

Figure 2. Inter-class distance and intra-class distance

In this paper, we propose AMS-CL by extending AM-softmax loss with center loss [5], which takes into account both inter-class distance and intra-class distance jointly. More specifically, AM-softmax is used to combine with a CNN model instead of using the conventional softmax, which aims to enlarge the inter-class distance. Simultaneously, center loss is introduced to play a role of constraining the intra-class data to be close to their class centers as much as possible, aiming to reduce the intra-class distance. As a result, both inter-class distance and intra-class distance are leveraged to facilitate face recognition. Extensive experiments demonstrate that taking into account both inter-class and intra-class relations improves the performance of face recognition significantly. The main contributions are summarized as follows:

- We propose the AMS-CL model to leverage the influence of intra-class distance and inter-class distance for face recognition. AMS-CL takes into account the spatial distribution of both intra-class and inter-class jointly, benefiting to the improvement of the performance on face recognition.
- We conduct extensive experiments on a public and real-world dataset. The experimental results show the effectiveness of our AMS-CL, which outperforms the approaches using softmax loss, AM-softmax loss, and center loss, etc.

2. RELEATED WORK

Face recognition has reached high performance in academia, benefiting from the convolution networks in recent years. DeepFace [7] adopts classic softmax loss to optimize the weights of CNN, and achieves accuracy of 97.35% on Labeled Faces in the Wild (LFW) [6], which is close to the human level. Sun et al. proposed DeepID3 [20] to ensemble VGG [21] and GoogleLeNet [22], which achieved accuracy of 97.45% on LFW. Google proposed FaceNet [1] that replaces softmax loss with triple loss, which minimizes the distance between an anchor and a positive sample, and maximizes the distance between an anchor and a negative sample. The method achieves an accuracy of 99.63% on LFW. Liu et al. [23] proposed a two-step learning method for face recognition, which combined multi-patch deep CNN with metric learning. They obtained an accuracy of 99.77% on LFW by training a large-scale dataset consisting of 1.2 million images for 18,000 persons. Data augmentation for face recognition [24] have been proposed and achieved an accuracy of 98.07% on LFW. In [5], the authors applied a center loss to restrict the data clustering with a center of each class. A new softmax loss function called L-softmax was proposed in [3], which optimizes cosine angles of feature and classifier directly, and achieves accuracy of 98.71% on LFW by pretraining on WebFace [8] dataset. The aforementioned approaches achieves quite high performance on LFW dataset. However, their performance largely depends on the scale of the datasets used for pre-training. In addition, the pre-training on the large-scale datasets are quite time-consuming and critical to hardwares, making them cannot be deployed on limited hardwares like personal computers. In addition, if deployed on personal or regular computers by using small-scale datasets for pre-training, their performance will decease dramatically.

In addition, Liu et al. proposed to extend A-softmax loss [2] with two restrictions ($\|W\|=1$ and $b=0$) based on L-softmax, and mapped the normalized weights with features into sphere. A-softmax achieves accuracy of 72.759% on MegaFace [9]. Wang et al. [4] introduced a novel additive angular margin for the softmax loss called AM-softmax. It has modified $cos\,\theta$ to $cos\,\theta - m$ to reduced computational complexity and accelerate the convergence of the model. As a result, the convergence speed of AM-softmax is faster than L-softmax loss. AM-softmax achieves an accuracy of 99.12% on LFW 6000 pairs. However, the generalization ability of the aforementioned models may be limited by using softmax loss. Furthermore, the models using triple loss, A-softmax, or L-softmax are hard to converge in the training stage, making them not easy-to-use in practice. In addition, the aforementioned models do not take into account inter-class distance and intra-class distance jointly, neglecting the spatial structures underlying inter-class and intra-class data samples.

3. METHODOLOGY

3.1 Overview of the Framework

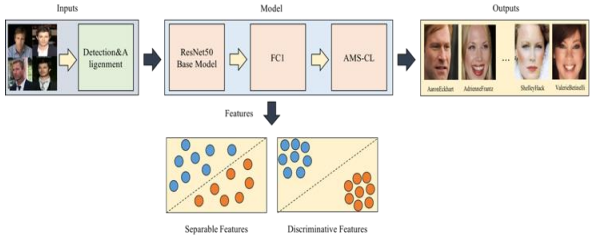


Figure 3. Overview of the framework

The overview of the framework is shown in Figure 3. Given the images, we conduct data alignment for the faces first. Furthermore, we build a fully-connected layer on the CNN network of ResNet50 [10] and use AM-softmax and center loss jointly in the last layer for face recognition. Finally, we obtain discriminative features from fc1 layer of the CNN model, which can be combined with the proposed AMS-CL for face recognition directly.

3.2 Data Processing

As the faces in natural scenes are not aligned and there exist background interference, we first detect the faces and their benchmark points in images by using MTCNN [18], which guarantees accuracy and can be conducted in real time. Furthermore, we use affine transformations to align the faces according to five benchmark points (i.e., two eyes, nose, and mouth corners). If one image contains two or more faces, we select the one with the largest area of detected frame. For some images that the faces cannot be detected, we simply exclude them for training and testing. As the faces in natural scenes are in different scales, we resize them to 224×224 RGB images as the input of CNN models. By convention, each pixel (in [0, 255]) in RGB images is normalized by subtracting 127.5 and divided by 128.

In particular, we do some data cleaning work as shown in Figure 4, as there may be a few negative samples in classes suffering from potential interference factors in face images of different datasets. For instance, sometimes a picture of a woman may appear in a class of man's images. As a result, we train a CNN models which are share the same weights of the layers on the face images to extract discriminative features, and calculate the cosine distance between the first face image in a class and the other classes. We assert the first face image of each class as a positive sample.

Furthermore, we choose a threshold to judge whether to save an image or not.

3.3 AM-Softmax with Center Loss

We propose the AMS-CL model by taking into account both intra-class distance and inter-class distance jointly. Our AMS-CL model consists of three components, i.e., convolution neural network, AM-softmax loss, and regularization term for center loss. As a result, we introduce the three components as follows, respectively.

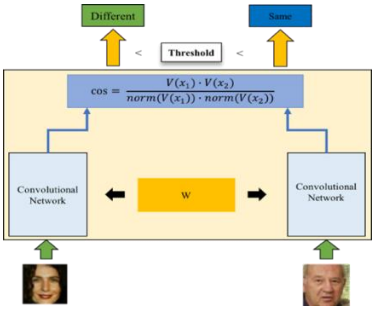


Figure 4. Data cleaning based on the cosine distance of features.

Convolution Neural Network. In recent years, convolutional neural networks (CNN) have made great progress in the field of computer vision. Extracting facial features using the CNN [11] network model is an important and fundamental component for face recognition. In particular, we adopt the ResNet50 model as the basic CNN structure in our work. ResNet model proposed a method called identity mappings, which can address the problems of gradient vanishing and extract advanced features from images. We keep the network structure consistent with [11], and replace the classifier adopted by imagenet [25] with a classifier for face recognition.

AM-softmax Loss. AM-softmax imports an angular margin into the softmax loss. The margin is formulated via $\cos\theta - m$, which is more simple and stable than [2], benefiting to the improvement of performance. The formulation is defined as follows:

$$L_{AM-S} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s.(cos\theta_{y_i}-m)}}{e^{s.(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c} e^{s.cos\theta_j}} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s.(W_j^T f_i-m)}}{e^{s.(W_j^T f_i-m)}+\sum_{j=1,j\neq y_i}^{c} e^{s.W_j^T f_i}} \quad (1)$$

where y is the label of the target data; y_i is expressed as the i-th category; f_i is the output of CNN layers; W is the weights; b is the bias in CNN layers. We initialize b as zero for simplicity. After normalizing weights and output of feature layer, the $\cos\theta$ is equal to $W \cdot f_i$. The symbol m is a value of margin between inter-class, and s is a scale to make the convergence stability. AM-softmax obtains a significant increase on the inter-class distance, and converges fast while training.

Center Loss. Center loss [5] is a regularization term, which can be combined with softmax loss directly for face recognition. Center loss aims to minimize the intra-class variations while keeping the features of different classes separable. Center loss constrains the distributions of the data in each class to be close to the center of the class as follows:

$$L_c = \frac{1}{2}\sum_{i=1}^{m}\|x_i - c_{y_i}\|_2^2 \quad (2)$$

where c_{y_i} is the center for each class.

AM-Softmax with Center Loss (AMS-CL). Consequently, we propose the AMS-CL by extending AM-softmax with center loss, in order to increase the inter-class distance and decrease the intra-class distance jointly. The formulation is defined as follows:

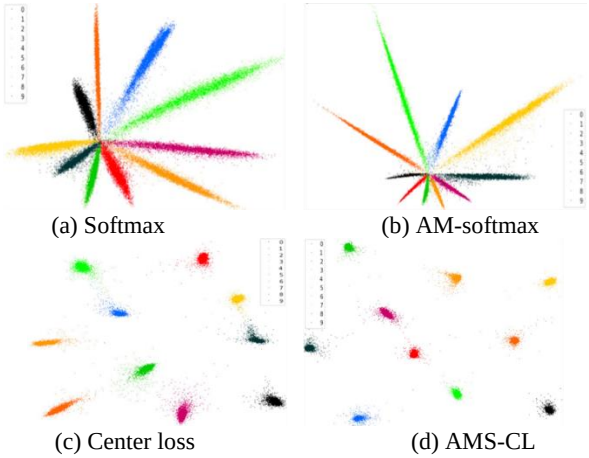
$$L_{our} = L_{AM-S} + \lambda L_c = -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s.(W_{y_i}^T f_i-m)}}{e^{s.(W_{y_i}^T f_i-m)}+\sum_{j=1,j\neq y_i}^{c} e^{s.W_{y_i}^T f_i}} + \frac{\lambda}{2}\|f_i - c_{y_i}\|_2^2 \quad (3)$$

The mathematic symbols are the same as the previous ones, and λ controls the impact of the regularization term.

For the optimizations of AMS-CL, gradient decent is adopted by following the back propagation algorithm [5].

3.4 Examples of the Intuition of AMS-CL

For illustration purpose, we give a toy example to show the intuition of AMS-CL as shown in Figure 5. We train four handwriting recognition models on the MNIST dataset [19] consisting of ten numbers by combing a CNN network with softmax loss, AM-softmax loss, center loss, and AMS-CL, respectively. Furthermore, we visualize the hidden data representation achieved by the four recognition models for the data samples of the ten numbers (i.e., class labels), respectively. From the figures, we have some observations. 1). By comparing Figure 5-a and Figure 5-b, we can see that AM-softmax has a larger margin of inter-class, and the data is more compact for each class than softmax. The experimental results demonstrate the significance of considering the intra-class distance by AM-softmax. 2). As shown in Figure 5-c, the inter-class distance among the different numbers (i.e., classes) are quite far than the the inter-class distances in Figure 5-a and Figure 5-b. As a result, the data samples can be classified more easily by using center loss. The experimental results demonstrate the significance of center loss for classification. 3). From Figure 5-d, we can see that the distance of inter-class is larger than AM-softmax and the intra-class distance is smaller than center loss at the same time, indicating the significance of the proposed AMS-CL. As a result, our AMS-CL takes the advantages of both AM-softmax for modeling intra-class distance and center loss for taking into account inter-class distance simultaneously.



(a) Softmax (b) AM-softmax

(c) Center loss (d) AMS-CL

Figure 5. Intuitive comparisons on intra-class distance and inter-class distance of the data representations achieved by the models.

4. EXPERIMENTS

4.1 Datasets

1) **FaceScrub** [12]. The dataset contains 70,053 images of 264 men and 265 women after data cleaning. Each class contains 132 images

on average. The minimum class contains 32 images, and the maximum class contains 255 images. Considering the imbalance of the dataset, we use the oversampling method in the training step. A few examples of the faces in the dataset are shown in Figure 6. In particular, the dataset is not quite large in terms of scale. We used it for pre-training in our experiments, aiming to explore the deployment of face recognition systems on regular computers as discussed in Section 2. The impact of the dataset size for pertaining is evaluated subsequently.

2) **Labeled Faces in the Wild (LFW).** The face photographs are collected for unconstrained face recognition. The dataset contains more than 13,000 images of faces collected from the Web with large variations in pose, expression, and illumination condition. Each face has been labeled with the name of the person pictured. There are 1,680 persons have more than two photos in the dataset. The test set contains 6000 face pairs in our experiments, and a few examples of the faces in the dataset are shown in Figure.7.

(a) Examples of an image containing one face

(b) Examples of an image containing two or more faces

Figure 6. Examples of FaceScrub dataset

(a) A pair of same person

(b) A pair of different person

Figure 7. Examples of LFW pairs

4.2 Baselines

The baselines include four models trained by using different loss functions, i.e., softmax loss alone, AM-softmax [4] alone (AMS), softmax joint with center loss [5] (CL-S), and the proposed AM-softmax with center loss (AMS-CL). For fair competitions, we adopt the same CNN network structure, i.e., ResNet50 for all the approaches with different loss functions. For implementations, we design the CNN model by using the Keras library with TensorFlow [13]. We split our dataset into a training set (90% of the data) and validation set (the rest 10%), and choose random crop and horizontal flip to augment data. The batch size of the CNN model is set as 64, and the learning rate starts from 0.1, which is divided by 10 at 20 epochs and 30 epochs. We choose the stochastic gradient descent (SGD) for optimization. The performance of the approaches is evaluated on the metric of accuracy.

4.3 Impact of the Parameter

There are three parameters in our AMS-CL, where m is the margin value between each class, s accelerates and stabilizes the optimizations, and λ is used to balance AM- softmax with center loss. According to the settings in paper [2], we set $s = 30$ and $m = 0.35$. The parameter λ dominates the intra-class variations and affects all the data distribution in spatial space in AMS-CL. Figure 8 shows impact of λ on the performance of AMS-CL on LFW. From the figure, we can observe that λ are critical to the performance of AMS-CL. In particular, our AMS-CL achieves the best performance on LFW when λ is 0.01.

4.4 Performance of AMS-CL on the LFW Dataset

We compare our model with other baselines on the LFW dataset in Figure 9. From the figure, we have the following observations.

1). AM-S outperforms the softmax model, indicating the significance of considering inter-class distance.

2). CL-S outperforms the softmax model, indicating the significance of considering intra-class distance by the regularization term of center loss.

3). Our proposed AMS-CL outperforms the baselines, indicating the significance of taking into account inter-class distance and intra-class distance jointly. The experimental results demonstrate the effectiveness and significance of taking into account the spatial distribution of both intra-class and inter-class jointly for face recognition tasks.

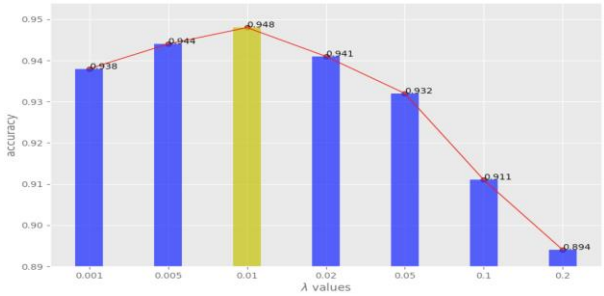


Figure 8. Impact of λ on the LFW dataset

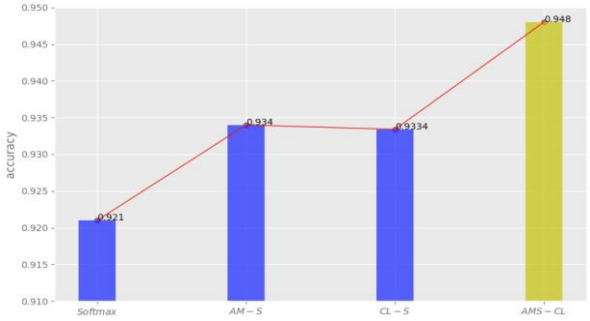


Figure 9. Performance of the approaches on the LFW dataset

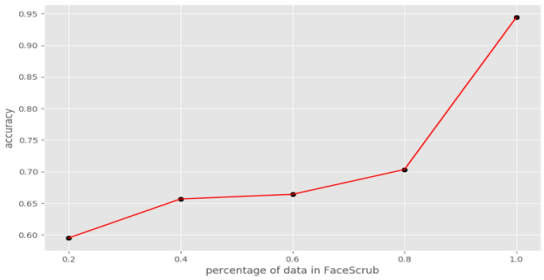


Figure 10. Impact of the percentage of data in FaceScrub for pre-training (AMS-CL)

4.5 Evaluation of Dataset Size for Pre-training

As discussed in Section 2, the dataset size for pre-training is critical to the performance of face recognition. As shown in Figure 10, the performance of AMS-CL increases dramatically with more data for pertaining. Therefore, it is expected that the performance of our AMS-CL can be improved further by using large-scale datasets for pre-training.

4.6 Performance of classifiers combined with AMS-CL

For face recognition, the features obtained for the testing images and labelled images are usually used directly for matching. Alternatively, we can also extract the features obtained by the deep learning models, and further train classifiers for face recognition. As a result, we further investigate quite a few classifiers on the features for face classifications, which can evaluate the effectiveness of the features obtained by the face recognition models in turn. Therefore, we implement a number of popular classifiers, including XGBoost [14], SVM, Random Forest [15], GBDT [16], and KNN [17], and report their performance on the current task as shown in Table 1. From the table, we can observe that using the features obtained by AMS-CL can achieve competitive performance for most of the classifiers. The experimental results demonstrate the effectiveness of the features obtained by our AMS-CL.

Table 1. Performance of the Classifiers on LFW

Model	XGBoost	SVM	Random Forest	GBDT	KNN
Softmax	0.918	0.9203	0.8827	0.9160	0.8815
AM-S	0.9002	0.9345	0.8995	0.8995	0.8985
CL-S	0.9153	0.9338	**0.9007**	0.9145	0.8995
AMS-CL	**0.9242**	**0.9493**	0.9005	**0.9230**	**0.9055**

5. CONCLUSION

In this paper, we propose the AMS-CL model by extending additive margin softmax with the center loss, considering inter-class distance and intra-class distance jointly. The proposed AMS-CL increases the inter-class distance and decreases the intra-class distance, which obtains more discriminative features for face recognition. In addition, we compared our AMS-CL with quite a few loss functions being used in face recognition on a real-world dataset. The experimental results show the effectiveness of the features obtained by AMS-CL, giving significant improvement on the performance.

6. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.61703109, No.91748107), and the Guangdong Innovative Research Team Program (No.2014ZT05G157).

7. REFERENCES

[1] Schroff, F., Kalenichenko, D., Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 815-823.

[2] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 212-220.

[3] Liu, W., Wen, Y., Yu, Z., Yang, M. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In International Conference on Machine Learning. pp. 507-516.

[4] Wang, F., Liu, W., Liu, H., Cheng, J. 2018. Additive Margin Softmax for Face Verification. In: IEEE Signal Processing Letters. pp. 926-930.

[5] Wen, Y., Zhang, K., Li, Z., Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In European Conference on Computer Vision. pp. 499-515.

[6] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst.

[7] Taigman,Y.,Yang,M.,Ranzato,M.,Wolf,L.2014.Deepface:Closingthegaptohuman- level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1701-1708.

[8] Yi, D., Lei, Z., Liao, S., Li, S.Z. 2014. Learning face representation from scratch. arXiv preprint arXiv:1411.7923.

[9] Miller, D., Kemelmacher-Shlizerman, I., Seitz, S.M. 2015. Megaface: A million faces for recognition at scale. arXiv preprint arXiv:1505.02108.

[10] He, K., Zhang, X., Ren, S., Sun, J. 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770-778.

[11] Sun, Y., Wang, X., Tang, X. 2014. Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1891-1898.

[12] Ng, H.W., Winkler, S. 2014. A data-driven approach to cleaning large face datasets. In: Image Processing (ICIP), 2014 IEEE International Conference on. pp. 343-347.

[13] Abadi, M., Agarwal, A., Barham, P. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv Preprint arXiv:1603.04467.

[14] Chen, T., Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785-794.

[15] Breiman, L. 2001. Random forests. Machine learning, 45(1), 5-32.

[16] Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[17] Cover, T., Hart, P. 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[18] Zhang, K., Zhang, Z., Li, Z., Qiao, Y. 2016. Joint face detection and alignment using multi-task cascaded convolutional networks. arXiv preprint arXiv:1604.02878.

[19] LeCun, Y., Cortes, C., Burges, C.J. 1998. The mnist database of handwritten digits.

[20] Sun, Y., Liang, D., Wang, X., Tang, X. 2015. Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873.

[21] Simonyan, K., Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[22] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1-9.

[23] Liu, J., Deng, Y., Bai, T., Wei, Z., Huang, C. 2015. Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint arXiv:1506.07310.

[24] Masi, I., Tran, A., T., Hassner, T., Leksut, J. T., Medioni, G. 2016. Do we really need to collect millions of faces for effective face recognition? In European Conference on Computer Vision. pp. 579-596.

[25] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.. pp. 248-255.

A New Method for Stroke Order Recognition of Handwritten Chinese Characters

Junjian Tang and Jun Guo
Computer Center
East China Normal University
3663 Zhong Shan Rd. N., Shanghai,
China
E-mail: jguo@cc.ecnu.edu.cn

ABSTRACT

Correctly writing Chinese characters, which are composed of strokes, is an essential education content in the primary and secondary school in China. Automatically recognizing the strokes order is very useful to help teachers and students find the mistakes occurred in the handwriting process. In this paper, we propose a new method for recognizing the strokes order. We collect plenty of images of each stroke in the Chinese character writing process from the handwriting board, and use convolutional neural network (CNN) to build the classification model. To deal with the characters that are easy to be written in a wrong order, we add connection points in the case of the two connected adjacent strokes, and connection lines in the unconnected case. Thus, the information of the strokes order can be described more distinctly. Using this kind of extracted features, we can recognize these fallible characters more effectively. In the experiments, the real Chinese characters in the textbooks of primary school are used, and the results show the feasibility and efficiency of our proposed method.

CCS Concepts

Computing methodologies ~ Computer vision

Keywords

Chinese character strokes order; convolutional neural network; classification; connection points; connection lines

1. INTRODUCTION

The education in Chinese characters writing is an important part of Chinese primary and secondary school students' education. The writing education of Chinese characters is mainly aimed at correctly writing Chinese characters, including writing correct strokes which constitute Chinese characters and correct writing order of strokes. Every Chinese character has a standard strokes order for writing, which is useful for the dissemination of Chinese.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301514

From primary school, students need to learn how to write Chinese characters through a lot of practice. The usual method is for students to practice with paper and pencil. Since the picture is 2-D, it is difficult to determine whether the stroke order of the Chinese characters is correct. With the popularity of smart phones, tablet computers and other electronic devices, we can easily record the writing process of Chinese characters. This greatly facilitates the Chinese character writing education of primary and secondary school students.

On-line handwritten Chinese character recognition is a very important research direction in pattern recognition. Many works have done in this area. While in stroke order recognition, some efforts are eminent.

Tan proposed a rule-based algorithm which is capable of identifying the typical writing errors in on-line handwritten Chinese character verification [1]. Chen *et al.* proposed an algorithm to recognize handwritten Chinese character which takes stroke segment as pattern primitive to represent the structure information of Chinese character [2]. Hu *et al.* built a distance education application of a Chinese handwriting education system that can automatically check the handwriting errors, such as the stroke production errors, stroke sequence error and stroke relationship error [3]. The system is based on attributed relational graph. Bai and Qiao established a mechanism of error correction for handwritten Chinese characters based on a new Chinese character structure description method [4]. Li *et al.* proposed a method to recognize stroke orders of on-line handwritten Chinese characters based on analyzing both spatial and temporal information [5]. This method adopted hidden markov model (HMM) in temporal information analysis.

This paper provides new insights into the verification of on-line handwritten Chinese character stroke order. During Chinese character handwriting education, students are often required to write strokes in neatly. We use handwriting pad to record the students' handwriting. The recognition is divided into two steps. First step is to identify the strokes. Recognizing the order of the strokes is the second. For the recognition of the stroke, a customary method is presented to match the shapes of the stroke. Convolutional neural network (CNN) is adopted in the stroke recognition. The spatial topological relations between adjacent strokes are applied in the recognition of strokes' order. In general, as long as one stroke is inconsistent with the standard stroke order, it can be determined that the entire character is written incorrectly. But there are some annoying situations that two characters consist of the same strokes and the orders of these strokes are also uniform. Considering for these situations, if the adjacent strokes are connected, we use connection point which is the meeting of two strokes to present the spatial topological relation. On the other

hand, we use connection line which is formed by connecting two strokes end to end to present the spatial topological relation. Applying the connection point and connection line to present the spatial topological relation between strokes prove useful in the verification of on-line handwritten Chinese character stroke order.

This paper has been organized in the following way. Section 2 shows the background of this method. Section 3 proposes a verification method for the order of on-line handwritten Chinese character strokes. Section 4 shows experimental results, and Section 5 is a conclusion.

2. BACKGROUND

In Chinese primary school textbooks, the strokes of Chinese characters have been divided into 28 categories. The combination of these 28 categories of strokes constitutes most Chinese characters. Fig. 1 shows all 28 categories of strokes.

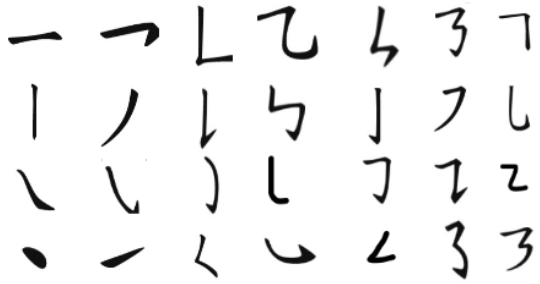


Figure 1. 28 categories of strokes

Due to the randomness of Chinese characters handwriting, the shapes of many strokes are similar. Fig. 2 shows two strokes which have similar shapes.

Figure 2. Similar strokes

Generally speaking, strokes with similar appearance will not appear continuously at the same time, so we can make a simple classification of many similar strokes.

The strokes are divided into 14 categories according to their shapes. Those 14 categories of strokes are shown in Fig. 3.

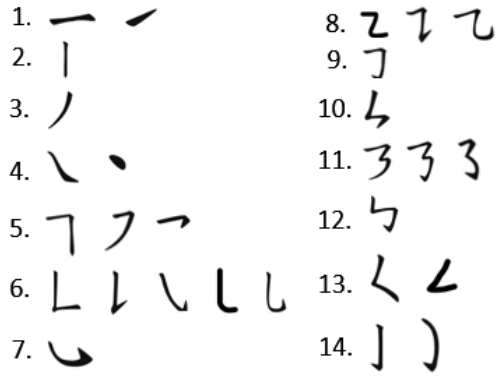


Figure 3. 14 categories of strokes

We use the handwriting pad to collect Chinese character handwriting. The hand-writing pad can store the image of each stroke and record the track of each stroke. These information have been stored in a XML file. The stroke images collected by the handwriting pad are the marks left on the handwriting pad each time the pen is placed and lifted. Take the center of the stroke as the center of the image, and extract the image of this stroke. The size of the stroke picture is 25×25. These images are used to recognize strokes. At the same time, the horizontal and vertical coordinates of the handwriting on the pad are saved in a XML file.

3. PROPOSED METHOD

According to the background mentioned above, because CNN has a good application in the field of handwritten character recognition, we choose CNN to recognize strokes. At the same time, the handwriting is used to inspect the correctness of the handwriting process. As for the correctness of stroke order, Chinese characters generally consist of many different strokes. Therefore, what we need is the correct stroke order and then compare the standard order with the input. For many Chinese characters with the same strokes and the same stroke orders, we apply connection point and connection line to check the stroke order. For connected strokes, we use the method of connecting point to determine the order of strokes so as to distinguish different Chinese characters. For unconnected strokes, we use the method of connecting lines to determine the order of strokes so as to ensure that the writing order is correct. Taking these spatial information into consideration makes the recognition more accurate.

3.1 Network for the Verification of Handwritten Chinese Character Strokes

CNN is a feedforward neural network that extract local features by convolving input with a group of kernel filters [6] [7]. It automatically extracts the most effective features during training. Therefore, it has a wide range of applications in the field of image recognition. it can avoid complicated feature engineering without extracting features, thus reducing a lot of repetitive and tedious data preprocessing work that is necessary when using traditional algorithms such as SVM.

Typical CNN consists of convolution layer, activation function layer, pool layer and full connection layer.

LeNet-5 is a well-known CNN that was first published in 1998 which was designed for handwritten and machine-printed character recognition [7]. With the development of deep learning, many new network model like GoogleNet, AlexNet and others have emerged. For the recognition of simple images such as strokes, we can well extract the features in the images by using a neural network that has fewer layers. Considering that we need to get the order of stroke in real time, we choose LeNet-5, which has fewer layers of network but is widely used in handwritten character recognition.

The whole network that has been used to recognize the handwritten Chinese character strokes consists of 6 layers. The input size is 32×32, followed by an alternating convolution layer and a sampling layer. The first volume layer includes 6 feature maps, the size of 28×28, the field size is 5×5, the activation functions is ReLu. And the next sampling layer also contains 6 feature maps, each feature map size is 14×14, the field size is 2×2, the max pooling is used to reduce the number of parameters within the model; second layer volume contains 16 feature maps, the size of 10×10, the neighborhood size is

5×5 and behind the sampling layer contains 16 feature maps, the size of 5×5, the neighborhood size is 2×2; the last is 3 fully connected layers, number of neurons were 240, 14 and 14. In the last fully connected layers, the softmax is used to classify the images. The whole network is shown in the Fig. 4.

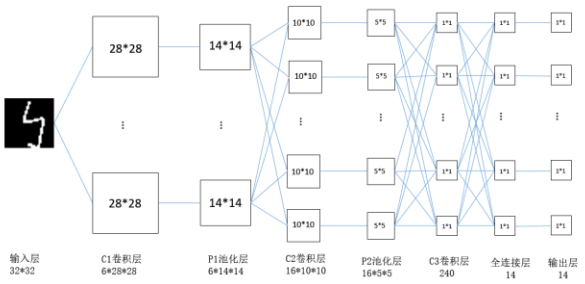


Figure 4. The whole network

3.2 The Correctness of Stroke Writing Process

The handwriting of a certain stroke has a specific trace. As shown in the Fig. 5, the correct handwriting trace of horizontal is from left to right, the correct handwriting trace of vertical is from top to bottom. For complicated strokes, the handwriting trace is relatively settled and is not easy to make mistakes. While for some simple strokes, it is easy to make mistakes because of its simple handwriting trace. Using the handwriting pad, we can get the horizontal and vertical coordinates of the writing trace of the pen, the writing speed, and the information of lifting and lowering the pen. The handwriting of strokes generally has a directional feature. This direction feature is the direction of the line connecting the starting point and the ending point of the stroke. Using this directional feature, we can know a trend of stroke writing, so it can well judge whether the stroke handwriting process is correct or not.

Figure 5. The correct handwriting trace

3.3 Connection Points

When learning to write Chinese characters, there are always some Chinese characters that make it very easy for us to make mistakes. They have same strokes and same orders of strokes, such as the two Chinese characters shown in Fig. 6 (a) and Fig. 6 (b).

(a)

(b)

Figure 6. Characters with same strokes and same strokes orders

Those characters with the same strokes and the same strokes orders bring troubles to the recognition of handwritten Chinese characters, and at the same time, they also bring troubles to the recognition of Chinese character strokes. For these Chinese characters, we propose a method of connecting points to identify these handwritten Chinese characters. For example, the correct stroke order serial of character in the left of Fig. 6 (b) is shown in the Fig. 6 (b), because of the wrong writing habits many people write the stroke 1 first and then write the stroke 3. With the information of connection point, we can detect this error easily.

On the handwriting pad, a stroke is recorded as a series of point sets. The positional relationship between the two point sets and their intersection points can reflect the order information between the two strokes. The problem that which one of the point sets is chosen to determine the positional relationship with the intersection point requires specific analysis for specific situations, such the two characters in the Fig. 6 (a), and the intersection point can be compared with any one of the two strokes. The intersection of the left character is in the head of the first stroke, in the middle of the second stroke. However, the intersection of the right character is in the middle of the first stroke, in the head of the second stroke. So regardless of the first stroke and the second stroke, the order relationship between the two strokes can be judged. For the two characters in the figure, the intersection of the first and second strokes of the left character is in the middle of the first stroke, and in the middle of the second stroke. The point of intersection of the right character is in the head of the first stroke, and in the middle of the second stroke. If we choose the second stroke and the intersection to determine the positional relationship, we cannot distinguish the difference between the two words. Therefore, we need to select the first stroke to make judgements.

After determining which stroke to use to determine the positional relation with the intersection, we need to assess the positional relation. We want to find the intersection of two point sets and use this to assess the positional relation. We describe these two strokes with two sets of points $A = \{(x_1, y_1), (x_2, y_2) \ldots\}, B = \{(x'_1, y'_1), (x'_2, y'_2) \ldots\}$, then the intersection of these two strokes can be considered as one of the two points that are closest to each other in the two point sets, then we choose one point which belongs to the second point sets as intersection point.

The algorithm for determining the intersection coordinates is shown in Fig. 7 as below.

Algorithm 1: Find the intersection point

Step 1: initialize $dis\_min = INT\_MAX$.
Step 2: For each $\boldsymbol{a}_q \in A$ *and for each* $\boldsymbol{b}_p \in B$, *if* $distance(\boldsymbol{a}_q, \boldsymbol{b}_p) < dis\_min$, *then let* $dis\_min = distance(\boldsymbol{a}_q, \boldsymbol{b}_p)$ *and set* $\boldsymbol{b}_p$ *as Intersection Point.*

Figure 7. Find the intersection point

After determining the coordinates of the intersection point, we need to judge whether the intersection point is at the head, middle or tail of the second stroke. We use the following algorithm to find the position of the intersection point. P_0 is the intersection point, P_1 is the head point in the second stroke, P_2 is the tail point in the second stroke. If $distance(P_1, P_0) < distance(p_2, p_0)$ *and* $(distance(p_2, p_0) - distance(p_1, p_0))/distance(p_1, p_2) > 0.7$

, then the intersection point is in the head of the second stroke. And if $distance(P_1, P_0) < distance(p_2, p_0)\ and\ (distance(p_2, p_0) - distance(p_1, p_0))/distance(p_1, p_2) > 0.7$, the intersection point is in the tail of the second stroke. If the situation is out of the situation mentioned above, the intersection point is in the middle of the second stroke

Figure 8. The connection lines

3.4 Connection Lines

There is also an error-prone situation in writing Chinese characters, in which two strokes are the same strokes and the writing order is reversed. This error is very difficult for the computer to judge, because the two strokes are the same, we cannot judge which stroke was written first. In this case, we judge the order of the two strokes by the direction of connecting line of the two same strokes.

The connection line is a ray formed by connecting two strokes end to end. The direction of the ray is from the tail of the first stroke to the head of the second stroke. As shown in the Fig. 8. We match this direction with the standard eight directions in the Fig. 9 get the direction number, and then compare it with the standard to judge whether the stroke order is correct.

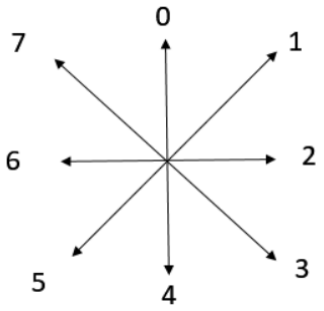


Figure 9. The 8 directions

The whole process of our method is shown in Fig. 10. We collect stoke images from the handwriting pad and then input the images into CNN. After the calculation of the CNN, we recognize the strokes. With the help of connection lines, connection points and direction features, we recognize the handwritten Chinese character strokes order in the end.

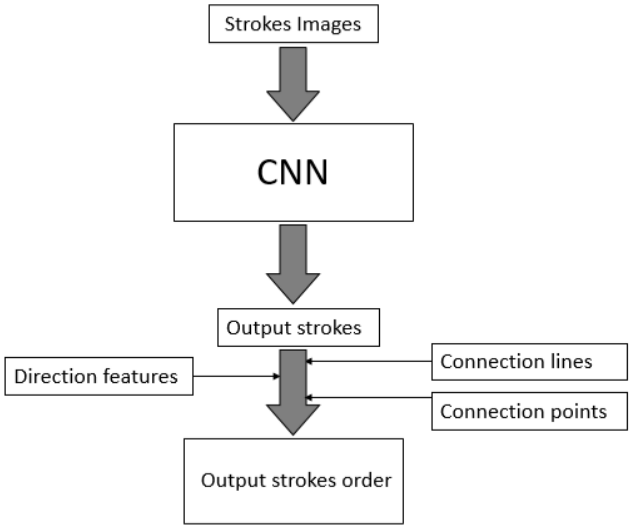


Figure 10. The whole process

4. EXPERIMENTAL RESULTS AND DISCUSSION

With the help of the handwriting pad, we collected 28 kinds of stroke data handwritten by different writers, 300 images for each stroke, 8400 images in total, and divided them into 14 categories. These images had been used as training data for CNN.

Since our method is aimed at the Chinese character writing education of primary and secondary school students, we cannot use public data sets and compare with other papers' results. We have designed three different experiments to verify the correctness of this method. The input device used in the experiment is a tablet with model NDE2-a5. We selected 400 characters from the Chinese vocabulary of the Chinese textbooks in first grade of primary school as the test samples. The following describes the three experiments in turn.

- Experiment 1: Characters whose strokes order can be recognized by standard stroke sequence

We selected 200 characters whose stroke order can be recognized by standard stroke sequence from the vocabulary. Every character had been written by different writers for 10 times. The total number of the characters is 2000.

- Experiment 2: Characters whose strokes order can be recognized by additional information

We selected 100 characters whose strokes order can be recognized by additional information from the vocabulary. Every character had been written by different writers for 10 times. The total number of the characters is 1000.

- Experiment 3: Mixed characters of Experiment 1 and Experiment 2

We selected 200 characters which was combined by 100 characters from Experiment 1 and 100 characters from Experiment 2. Every character had been written by different writers for 10 times. The total number of the characters is 2000.

The result is shown in Table 1.

Table 1. Results in different experiments

	Total Strokes Images	Wrong Recognition	Recognition Rate
Experiment 1	2000	56	97.20%
Experiment 2	1000	47	95.30%
Experiment 3	2000	70	96.50%

From the experimental results, we can see that this method can deal with the situations that the writing order of the same strokes was reversed and the writing order was wrong due to the wrong writing habits very well.

The mistakes in Experiment 1 are almost caused of the wrong recognition of strokes. The randomness of handwriting makes stokes hard to be recognized. As for the errors in Experiment 2, some of them have the same reason with the errors in Experiment 1. And irregularities in writing process are also responsible for the inaccurate verification. Irregularities in writing process lead to mistakes in judgement of connection points and connections, thus making mistakes in recognition of stroke order.

5. CONCLUSION

In this paper, we proposed a new method for recognizing the strokes order, which takes the spatial topological information of the adjacent strokes into consideration. This method is useful for helping primary students to practice writing Chinese characters. In the future, more strokes samples could be collected to make the CNN more efficient. And using other CNN model may be also a nice try. We will develop a smartphone application based on this method to help more people to learn the Chinese characters handwriting.

6. REFERENCES

[1] Chwee Keng Tan. An algorithm for online strokes verification of Chinese characters using discrete features. pages 339–344, 2002.

[2] Chen Zhi. A recognition algorithm of chinese character based on stroke segment and order. Journal of Hunan University, 2000.

[3] Zhihui Hu, Yun Xu, Liusheng Huang, and Howard Leung. A chinese handwriting education system with automatic error detection. Journal of Software, 4(2):101–107, 2009.

[4] Xiaodong Bai and Xiaojun Qiao. A method of chinese character shape representation and its application in the error correction for normative handwritten chinese characters. pages 1–7, 2017.

[5] Rongsha Li, Liangrui Peng, Endong Xun, and Nan Wei. A stroke order verification method for on-line handwritten chinese characters based on tempo-spatial consistency analysis. pages 999–1003, 2013.

[6] Yann Lecun, Bernhard E Boser, John S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Handwritten digit recognition with a back-propagation network. pages 396–404, 1990.

[7] Y. Lecun. Gradient-based learning applied to document recognition. Intelligent Signal Processing, pages 306–351, 2001.

Kick Recognition System Using Dual RGB-D Cameras

Sungjin Hong
Creative Contents Research Division
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
sjhong0117@etri.re.kr

Myung-Gyu Kim
Creative Contents Research Division
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
mgkim@etri.re.kr

Yejin Kim
School of Games
Hongik University
Sejong, Republic of Korea
yejkim@hongik.ac.kr

ABSTRACT
This paper introduces a kick recognition system based on human body detection using dual RGB-D cameras. Recently, the availability of RGB-D cameras makes it possible to get the human body joints that are informative for activity analysis. However, single camera-based approaches enforce frontal-oriented action due to the occlusion problem. Using dual RGB-D cameras and a smart sandbag, the proposed system detects major joints and recognizes various kick actions from a general user in real time. For each camera, our system detects salient body parts in a kick action such as a head and feet. A local detector trained with a supervised model is used for the head detection. The detected body parts are converted into a quadtree-structured graph model to detect feet using accumulative geodesic distance (AGD). To deal with the occlusion by the sandbag, our system compares the result of each camera and selects the most reliable one based on AGD. For the kick recognition, a finite-state machine is adopted to track and to segment continuous kick movements into different states. Considering a viewpoint change and a variable kick speed, fixed size descriptors are constructed from the interpolated action to recognize user kicks. We evaluated our system using various kick actions in taekwondo and achieved a high recognition rate of 92%.

CCS Concepts
• **Computing methodologies → Activity recognition and understanding.**

Keywords
Human kick recognition, RGB-D cameras, body joint detection, feature representation, sandbag experience system.

1. INTRODUCTION
Human pose estimation and human activity recognition (HAR) have been widely studied in the computer vision field. The recent advent of inexpensive RGB-D cameras makes it possible to get depth image data at a low cost and to detect body joints in real time. Owing to these advantages, the joint-based HAR has been applied to various real applications in human computer interface, healthcare, and sports.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301517

To detect a human body using a single RGB-D camera, Shtton *et al.* performed the pixel-level classification using a randomized forest tree and estimated body joints with a mean-shift algorithm [14]. Recently, a joint detection method based on convolutional neural network (CNN) has been proposed with deep learning methods [9,11]. These methods can estimate a human pose regardless of a user's body size; however, they require a large amount of training data. Plagemann *et al.* searched major joints using accumulative geodesic distance (AGD) and used the local detectors to classify the joints [13]. In their approach, the major joints such as a head, hands, and feet are located far from the center of the body. Hong and Kim [6] and Baak *et al.* [1] reduced the computation cost by constructing quadtree-structured graph model and using optimal graph search for real time performance. These AGD-based approaches are robust to a view-changing direction and requires less training data than the learning-based approaches. However, they are difficult to detect minor joints such as elbows, knees, and shoulders that are relatively close to the center of the body.

Recently, the single camera-based approaches have extended to adopt multiple cameras due to the occlusion problem [2,8,10]. These approaches have mainly focused on unifying a set of skeleton data detected at different viewpoints into a single skeleton. Baek and Kim [2] divided the human body into five components (e.g., torso, arms, legs) and selected a main camera for each component based on the tracked status of the joints. Kim *et al.* [8] proposed the different tracking methods for each joint to acquire the optimal skeleton including rotation of 360 degrees. These methods rely on a trained model assuming a frontal oriented posture, which requires a sophisticated detection algorithm, especially for the body parts captured from the side cameras.

In joint-based HAR, a pair-wise difference is computed to extract spatial-temporal descriptors. Yang and Tian [15] proposed a method called EigenJoint which combines static, dynamic, and offset features between frames. Kapsoura and Nikolaidis [7] used pair-wise angular information to recognize actions using the codebook approach. Pair-wise difference can effectively represent human movements; however, its accuracy decreases when all of body joints are used in HAR. Ofli *et al.* [12] and Chen *et al.* [3] proposed partial joint-based approaches by determining joint priorities. Olf *et al.* partitioned an action into segments and represented a sequence of the segments with the most informative joints. Like [15], Chen *et al.* determined partial joints by clustering and computing the difference of the joints. Using the partial joint-based approaches can reduce a computation cost and improve a detection accuracy.

This paper introduces a kick recognition system based on human body detection using RGB-D cameras. Using dual RGB-D cameras, we propose a sandbag experience system that detects major joints and recognizes various kick actions from a user in

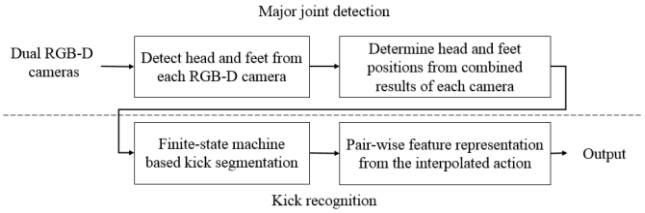


Figure 1. Overview of the proposed system.

real time. Our system detects salient body parts in a kick action such as a head and feet. A local detector trained with a supervised model is used for the head detection while the detected body parts are converted into a quadtree-structured graph model to detect feet using accumulative geodesic distance (AGD). To deal with the occlusion by a sandbag, our system compares the result of each camera and selects the most reliable one based on AGD. For the kick recognition, a finite-state machine is adopted to track and to segment continuous kick movements into a state. Considering a viewpoint change and a variable kick speed, fixed size descriptors are constructed from the interpolated action for a kick recognition. The overview of our system is shown in Figure 1.

The rest of this paper is organized as follows. Section 2 describes the major joint detection using dual RGB-D cameras. Section 3 presents a segmentation method and feature representation for kick recognition. Section 5 shows the experimental results, and Section 6 concludes this paper.

2. MAJOR JOINT DETECTION

2.1 Single camera process

Given color and depth images as inputs from a camera, background data are subtracted from the images to expedite the detection process based on AGD. The background data are defined as a depth image at the first frame. To reduce a computational cost, we utilized the approaches like [1,6] which construct foreground data into a quadtree-structured graph model and search for an optimal graph using Dijkstra algorithm. For the graph search, a starting point is initialized with a head as feet are located farthest away from the head. Without using local detectors, the feet can be searched by AGD while the head can be robustly detected by hog descriptors and support vector machine (SVM) which trains a model with a relatively small number of samples [5]. If the detection is failed, a mean-shift algorithm is used to track the head from a previous frame [4]. In our system, the head position is searched first, and then two farthest positions are searched from the head using AGD for the feet.

2.2 Dual camera process

To minimize occlusions by a sandbag and a user self-occlusion, dual cameras are adopted in our system. As a preprocessing step, the local coordinates of each camera should be calibrated into the same system (i.e., the world coordinate system). One camera is selected as a reference coordinate system, and a transform matrix is estimated by minimizing errors between the reference and other local coordinate, using a checkerboard as a calibration tool. To combine the result from each camera, two cases are considered: First, both cameras successfully detect the major joints. In this case, the results are closely placed in the world coordinate system if the difference between their results is less than 100mm. The major joints are estimated from either an averaged position or selecting one of the camera results. Next, if one camera fails to detect the major joints owing to the occlusion problems, the results are not closely located in the world coordinate system. In

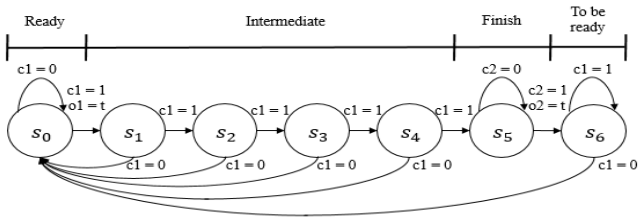


Figure 2. Finite-state machine for a kick action.

general, the properly detected joints have longer AGD than the incorrectly detected joints. We compare the sum of AGD of each camera to choose reliable one. The joint positions are determined by maximizing the sum of AGD as follows:

$$P^t = \arg\max_{J_c^t} \sum_{j \in J_c^t} AGD(j), \quad (1)$$

where J_c^t is a set of the detected joints in the c indexed camera at frame t, and P^t is the determined joint at frame t.

The left and the right side of the feet is distinguished from an initial pose which is assumed to stand and to look forward. At the initial pose, the left and the right side of the feet are labeled based on $+x$ axis. At time t, P^t are labeled to minimize a Euclidean distance from P^{t-1}. The set of labels is $\{HE, LF, RF\}$ for head, left foot and right foot, respectively.

A viewing direction of the torso is estimated to analyze the rotation of kicks. Given a set of 3D points, principal component analysis (PCA) can find the best fitting plane by minimizing the sum of square errors. The first and second principal components represent plane axis while the third one represents the direction of the plane. We assume that the torso lies within a radius 200mm from the center of the point cloud. The torso is analyzed with PCA, and the third principal component is set as the viewing direction. It is noteworthy that the principal components can be further used to align the reference coordinate system with a specific direction as shown in Section 3.2.

3. KICK RECOGNITION

3.1 Segmentation

Finite-state machine (FSM) defines a finite number of states and transition condition to control the behavior of objects according to certain rules. In our system, FSM is designed to determine start and end of kick actions by tracking the state based on rules.

Figure 2 shows FSM for a kick action that is divided into four phases: The first one is a ready action that feet stay on the floor (s_0). The second one is an intermediate action that one leg is on the floor while the other is lifting ($s_1 \sim s_4$). The third one a finishing action that the lifting leg is stretching in the air (s_5). The fourth one is a ready action from the last action (s_6). If all four phases are observed sequentially, this action is judged as a kick. As shown in Figure 2, $s_0 \sim s_4$ and s_6 perform a state transition according to input $c1$ while s_5 performs a state transition according to input $c2$. Output $o1$ and $o2$ are defined in state s_0 and s_1 when input $c1=1$ and $c2=1$, respectively. Here, $s_1 \sim s_4$ are designed to consider a continuous kick action at least 5 frames. Input $c1$ is defined as follows:

$$d1 = \left\| p^t_{y,i=LF} - p^t_{y,i=RF} \right\|_2,\ d2 = \left\| p^t_{(x,z),i=LF} - p^t_{(x,y),i=RF} \right\|_2$$

$$c1 = \begin{cases} 0, & if\ d1 < H \times r1 \vee d2 < H \times r2 \\ 1, & otherwise \end{cases} \quad (2)$$

where $p^t_{y,i}$ is the *y*-axis position of the joint with label i at time t, H is a user's height, and $r1$ and $r2$ are ratio of the height. In our system, $r1$ and $r2$ are set to 0.1 and 0.2, respectively. When $c1 = 1$, the state is shifted from s_0 to s_1, and it produces the output value, $o1$, which is set as the start time of the kick, $t1$.

Input $c2$ is defined as follows:

$$\hat{k} = \arg\max_{-\frac{w}{2} \le k \le \frac{w}{2}} \left\| p^{t+k}_{i=LF} - p^{t+k}_{i=RF} \right\|_2$$

$$c2 = \begin{cases} 0, & if\ \hat{k} \neq 0 \\ 1, & otherwise \end{cases}, \quad (3)$$

where w is the size of time window for the foot detection, the farthest position between left foot and right foot. In our system, w is set to 6. When $c2 = 1$, the state is shifted from s_5 to s_6, and it produces the output value, $o2$, which is set as end time of the kick, $t2$. At last, the state is shifted from s_5 to s_6 when a kick action is to be ready.

3.2 Feature representation

The feature descriptors are represented using major joints and the viewing direction of the torso coming from the dual camera process. These features use both invariant viewing direction and difference movement speed. To minimize the effect of variable viewing direction, the world coordinate system is aligned with a user's specific direction at $t1$. The action at $t1$ is standing that is a ready pose. In the dual camera process, the third principal component from the eigenvectors of the torso is used as the viewing direction. This vector is orthonormal to the first and second principal components that are represented approximately vertical and parallel to the ground, respectively. To normalize the viewing direction, *+y*, *+x*, and *+z* axis of the world coordinate system is set to the first, second, and third principal component, respectively. The origin of the system is set to the center of the torso.

The major joints are interpolated with B-spline curves and the viewing direction is interpolated linearly to handle difference movement speed. This interpolation ensures continuity and extraction of features at arbitrary time. In our system, uniform cubic B-spline curves are used to interpolate trajectories of the major joints. Given the joint positions, an interpolated curve Q can be defined as follows:

$$Q_{t1 \le j \le t2}(u) = \frac{1}{6} \begin{bmatrix} u^3 & u^2 & u^1 & u^0 \end{bmatrix} \begin{bmatrix} -1 & 3 & 1 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 0 & 3 & 0 \\ 1 & 4 & 1 & 0 \end{bmatrix} \begin{bmatrix} P^{i-3} \\ P^{i-2} \\ P^{i-1} \\ P^i \end{bmatrix}, \quad (4)$$

where P^i is a control point which is set from the joint position at frame i, and u is a normalized factor [0,1] for each of the curves.

The viewing direction is linearly interpolated that is defined as follows:

$$\Theta_{t1 \le j \le t2}(u) = \begin{bmatrix} 1-u & u \end{bmatrix} \begin{bmatrix} \theta^{j-1} \\ \theta^j \end{bmatrix}, \quad (5)$$

where θ^j is the viewing direction of the torso at frame j, and u is a normalized value similar to Q. In addition, the interpolated trajectories and the viewing direction are normalized with a user's height and scaled by 180 degrees to reduce an intra class variation as follows:

$$NQ = \frac{Q}{H},\ N\Theta = \frac{\Theta}{180^\circ} \quad (6)$$

In our system, N number of pair-wise difference of NQ and $N\Theta$ are computed respectively. When F_j is a pair-wise joint position between NQ at frame $t1$ and at arbitrary time t', it can be estimated as follows:

$$F_j = \left\{ i^{t'}_l - j^{t1}_l : i, j \in NQ; l \in label; t' = t1 + \frac{|t2 - t1|}{N} \times k, 0 < k \le N \right\} \quad (7)$$

When F_c is a pair-wise angle between $N\Theta$ at frame $t1$ and at arbitrary time t', it can be estimated as follows:

$$F_c = \left\{ i^{t'} \cdot j^{t1} : i, j \in N\Theta; t1 < t' \le t2; t' = t1 + \frac{|t2 - t1|}{N} \times k, 0 < k \le N \right\} \quad (8)$$

In our system, $N = 30$ to construct fixed size descriptors. F_j and F_c are concatenated as $F = [F_j, F_c]$ that are used to train a model for kick recognition.

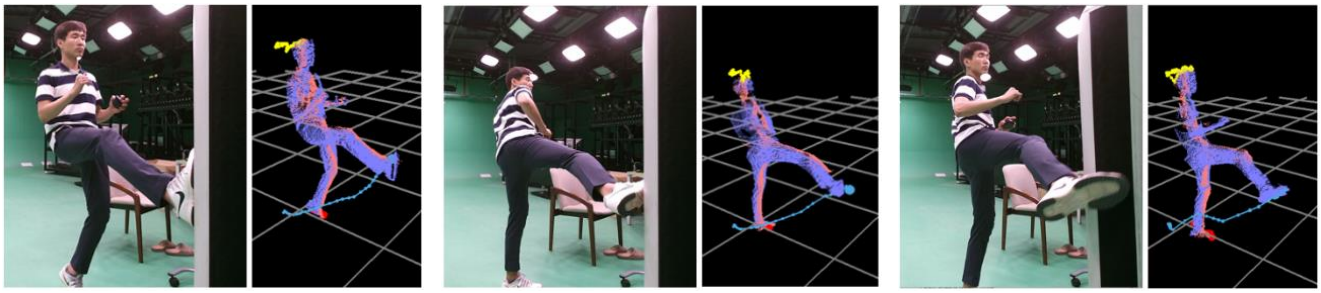
4. EXPERIMENTAL RESULTS

Figure 3 shows the sandbag experience system built for major joint detection and kick recognition. In the system, a user hits each side of the sandbag with various types of kicks and scores a higher point if the user's kick matches the type instructed by the system. For RGB-D cameras, two MS Kinects (version 2) are placed the right and left side of the sandbag. As shown in Figure 4, we evaluated the performance of our system using three types of taekwondo kick actions such as front, side, and round kick. Each kick is subdivided into the left and right side.

First, we tested the accuracy of major joint detection by performing the kick actions continuously without order. The sequences of the actions consist of 1434 frames from 60 samples. The major body parts were successfully detected with 97% for head, 95.05% for the right foot, and 95.33% for the left foot when the error threshold is set to 300mm. Most of the detection errors occur when the system falsely detects the sandbag as a body or the right and left side of feet from kick crossing actions.

Next, we tested the accuracy of recognition rate from a total of 607 samples that is composed of 420 training samples and 187 test samples. A model was trained with SVM using the features F. Table 1 summarizes the performance of this test. The system achieved the average rate of 92% for the test samples. However, the side kick accuracy is 10% lower than the averaged one due to the false detection of kick exchanges between the left and right side of the foot.

Figure 3. The sandbag experience system using our method.



(a) Front kick (b) Side kick (c) Round kick

Figure 4. RGB image and trajectories of the major joints for three types of kicks used in our experiments.

Table 1. The accuracy of recognition rate for taekwondo kick actions.

Motion types		Front kick		Side kick		Round kick	
		Left	Right	Left	Right	Left	Right
Front kick	Left	0.97	0	0.03	0	0	0
	Right	0	1.00	0	0	0	0
Side kick	Left	0	0	0.82	0.15	0.03	0
	Right	0	0	0.10	0.87	0	0.03
Round kick	Left	0	0	0	0	1.00	0
	Right	0	0	0	0.07	0.07	0.86

5. CONCLUSION

In this paper, we introduced a kick recognition system based on human body detection using dual RGB-D cameras. The joint detection uses a local detector and AGD to find head and feet that are key body parts in kick activities. To solve the occlusion problems, we compare the sum of AGD of each camera to select reliable positions of the major joints. To recognize the user kick, a finite-state machine is designed to segment a continuous kick action into different states. Considering a viewpoint change and a variable kick speed, fixed size descriptors are constructed from the interpolated action to extract features at the arbitrary time. The experimental results demonstrated that our system achieves a high recognition rate of 92% for taekwondo kicks and can be applied to the sandbag experience system.

Current system is mainly designed to recognize kick actions in Taekwondo. We believe that hands are another key body part, which can be detected by using AGD started from the head and body part detectors. This way, the tracking and recognition method can be extended to recognize more diverse Taekwondo actions such as punches and defensive poses.

6. ACKNOWLEDGMENTS

This research project was supported by The Sports Promotion Fund of Seoul Olympic Sports Promotion Foundation from Ministry of Culture, Sports and Tourism (S072016122016) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017R1C1B5017000).

7. REFERENCES

[1] Baak, A. et al. A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera. 2011 IEEE *International Conference on Computer Vision*, (Nov. 2011) 1092-1099.

[2] Baek, S. and Kim, M. 2015. Dance Experience System Using Multiple Kinects. *International Journal of Future Computer and Communication*. 4, 1 (Feb. 2015), 45–49. DOI=https://doi.org/10.7763/IJFCC.2015.V4.353.

[3] Chen, H. et al. 2016. A novel hierarchical framework for human action recognition. *Pattern Recognition*. 55, (Jul. 2016), 148–159. DOI=https://doi.org/10.1016/j.patcog.2016.01.020.

[4] Comaniciu, D. et al. 2000. Real-time tracking of non-rigid objects using mean shift. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)* (Hilton Head Island, SC, USA, 2000), 142–149.

[5] Dalal, N. and Triggs, B. 2005. Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (San Diego, CA, USA, 2005), 886–893.

[6] Hong, S. and Kim, M. 2016. A Framework for Human Body Parts Detection in RGB-D Image. *Journal of Korea Multimedia Society*. 19, 12 (Dec. 2016), 1927–1935. DOI=https://doi.org/10.9717/kmms.2016.19.12.1927.

[7] Kapsouras, I. and Nikolaidis, N. 2014. Action recognition on motion capture data using a dynemes and forward differences representation. *Journal of Visual Communication and Image Representation*. 25, 6 (Aug. 2014), 1432–1445. DOI=https://doi.org/10.1016/j.jvcir.2014.04.007.

[8] Kim, Y. et al. 2017. Motion Capture of the Human Body Using Multiple Depth Sensors. *ETRI Journal*. 39, 2 (Apr. 2017), 181–190. DOI=https://doi.org/10.4218/etrij.17.2816.0045.

[9] Moon, G. et al. 2017. Holistic Planimetric prediction to Local Volumetric prediction for 3D Human Pose Estimation. *arXiv:1706.04758 [cs]*. (Jun. 2017).

[10] Moon, S. et al. 2016. Multiple Kinect Sensor Fusion for Human Skeleton Tracking Using Kalman Filtering. *International Journal of Advanced Robotic Systems*. 13, 2 (Mar. 2016), 65. DOI=https://doi.org/10.5772/62415.

[11] Nishi, K. and Miura, J. 2017. Generation of human depth images with body part labels for complex human pose recognition. *Pattern Recognition*. (Jun. 2017). DOI=https://doi.org/10.1016/j.patcog.2017.06.006.

[12] Ofli, F. et al. 2014. Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*. 25, 1 (Jan. 2014), 24–38. DOI=https://doi.org/10.1016/j.jvcir.2013.04.007.

[13] Plagemann, C. et al. 2010. Real-time identification and localization of body parts from depth images. *2010 IEEE International Conference on Robotics and Automation* (Anchorage, AK, May 2010), 3108–3113.

[14] Shotton, J. et al. 2011. Real-time human pose recognition in parts from single depth images. *Computer Vision and Pattern* Recognition. *CVPR 2011* (June. 2011), DOI=https://doi.org/10.1109/CVPR.2011.5995316.

[15] Yang, X. and Tian, Y.L. 2012. EigenJoints-based action recognition using Naive-Bayes-Nearest-Neighbor. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (Providence, RI, USA, Jun. 2012), 14–19.

LoID-EEC: Localizing and Identifying Early Esophageal Cancer Based on Deep Learning in Screening Chromoendoscopy

Xiaoxiao Du
School of information engineering, Zhengzhou University
Zhengzhou, China, 450000
dxxzzu@outlook.com

Ya Li
The First Affiliated Hospital of Zhengzhou University
Zhengzhou, China, 450000
liya0305@126.com

Jianning Yao
The First Affiliated Hospital of Zhengzhou University
Zhengzhou, China, 450000
jianningyao@163.com

Bing Chen
The First Affiliated Hospital of Zhengzhou University
Zhengzhou, China, 450000
chenbing_1121@163.com

Jiayou Song
School of information engineering, Zhengzhou University
Zhengzhou, China, 450000
iejysong@zzu.edu.cn

Xiaonan Yang
School of information engineering, Zhengzhou University
Zhengzhou, China, 450000
iexnyang@zzu.edu.cn

ABSTRACT
Esophageal cancer is one of the most common malignant tumors which responses for about 400,000 deaths each year. Early identifying lesions is critical for reducing esophageal cancer mortality and the overall esophageal cancer burden. However, identification of early esophageal cancerous lesions can be very challenging for clinicians owing to the mild clinical symptoms and lack of specificity of esophageal cancer. Consequently, precancer or subtle early neoplastic changes may not be evident, limiting the diagnostic accuracy. As a clinical assistance for early esophageal cancer identification, a deep learning framework referred to as the M-Deeplab model was proposed for the localization and recognition of esophageal mucosa lesion. The proposed M-Deeplab model was extended from the Deeplabv3+ model by employing an encoder-decoder structure for accuracy improvement. It achieves high-precision semantic segmentation for different staining degrees and different sizes of endoscopic images. The overall accuracy reaches 97.31% and the MIoU reaches 92.09%. Moreover, it takes only 0.05s to judge one image by the M-Deeplab model. The M-Deeplab model exhibits good performance both in accuracy and speed for early esophageal cancerous lesions identification, comparable to the experienced clinicians. As an assistance for the clinicians, the proposed model could possibly increase the early esophageal cancer diagnosis accuracy and decrease the misdiagnosis.

CCS Concepts
•Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems → Image segmentation

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301540

Keywords
Early Esophageal Cancer; Endoscopy; Deep Learning; Semantic Segmentation

1. INTRODUCTION
Esophageal cancer ranks as the sixth leading cause of cancer-related death worldwide, accounting for an estimated 400,000 deaths each year, especially in China and Japan [1]. Patients with early esophageal cancer convey a much higher 5-year survival rate of up almost 30% than those with advanced esophageal cancer [2]. Survival outcomes for early diagnosed esophageal cancer are superior to outcomes in those with advanced disease at diagnosis. Therefore, early identifying lesions is critical for reducing esophageal cancer mortality and the overall esophageal cancer burden [3].

Early esophageal cancer refers to the stage that tumors haven't grown beyond the second layer of the esophagus wall and there is no cancer in lymph nodes. Early esophageal cancer is diagnosed from neoplasia limited to the esophageal mucosa and superficial submucosa (invasion of submucosa <500 μm), without any lymphatic or vascular invasion [4]. Endoscopic screening techniques have been evolved extensively over the past few years and have been implemented in early esophageal cancer diagnosis [5,6]. However, endoscopy is still limited in the detection rate of early esophageal cancer currently. Early intramucosal neoplasia can appear as small erosions, flat mucosal lesions, or normal mucosa, making it difficult to observe pathological changes clearly under the naked eyes. This may cause misdiagnosis or missed diagnosis. In contrast, stains help most in the delineation of superficial neoplastic lesions. However, the judgment of endoscopic esophageal staining is susceptible to subjective factors of the operators, which affects the choice of biopsy site and the diagnosis rate of early esophageal cancer. Consequently, it is highly desirable to develop an unbiased method to improve the detection rate of early esophageal cancer and precancerous lesions.

Computer Aided Diagnosis (CAD), especially with the help of deep learning technology, has gradually become a research hotspot in the diagnosis field owing to its autonomous learning capability. CAD helps to reduce the work intensity of doctors and provides effective auxiliary diagnostic information for disease

diagnosis [7]. In the area of automatic diagnosis of fundus diseases, the automated diagnosis method developed by the Google team has exceeded the average level of human doctors in accuracy and speed [8]. Esteva et al. demonstrated classification of skin lesions using a single neural network, they developed an automatic algorithm to distinguish keratinocyte carcinomas from malignant melanomas with training end-to-end from images directly, reaching the level of human experts [9]. Na Niu et al. proposed a quantitative analysis approach in multi-color spaces for early esophageal cancer diagnosis [10]. This method analyzed the color space in a fixed-size rectangular frame, but still causing a certain error for the iodine-stained area with irregular characteristics. Wu Ye et al. calculated the average values of G (Green), U (Chroma) and V (Brightness) in three color spaces (RGB, YUV, HSV) on the basis of Niu's method [11]. However, the CAD assisted early esophageal cancer diagnostic is not accurate enough with respect to human experts. One of the main limitations is that these auxiliary diagnostic methods need to manually select features, confound the results with the bias of human choice. Recently, the convolutional neural network (CNN) has been proposed to automatically select features instead of previous ways of artificial ways. With numerous of successful applications in many fields [12], such as computer vision [13], speech recognition [14], natural language processing [15], games [16], and biology [17] etc., CNN has been proved highly effective in image analysis and exhibited high-potential for cancer lesions identification. However, the detection of lesions on early esophageal cancer images by deep learning techniques remains to be studied.

In this work, a deep learning method referred to the M-Deeplab model was proposed for early esophageal cancer lesions localization and recognition. This model was extended from the Deeplabv3+ mode (one type of CNN) by applying an encoder-decoder structure for accuracy improvement. This method gave satisfactory segmentation results with MIoU ups to 92% (pixel accuracy > 97%). Compared with traditional methods, this method not only identifies lesions but also locates lesions on images with high accuracy and high speed, greatly reducing the probability of misdiagnosis and missed diagnosis. These results pave the way of deep learning for early esophageal cancer diagnosis.

2. EXPERIMENTS

2.1 Image Acquisition

The early esophageal cancer image data is acquired from the chromoendoscopy. The changes of esophageal mucosa can be visually observed under endoscopy, and the tumor status can be evaluated. The nature, location, boundary, and extent of the lesion can be evaluated by staining and magnification. The suspected esophageal lesions of 110 patients were dyed by Lugo's iodine solution [18]. Material includes: Olympus EVIS-260 host, GIF-XQ260 gastroscope, Spray tube, 1.2% Lugols iodine solution (12g iodine and 24g potassium iodide, diluted to 1000mL with distilled water, stored in a refrigerator at 4℃, storage time not exceeding 1mo).

After completion of the informed consent process, screened participants were provided a local anesthetic (5mL of 1% lidocaine by mouth for 5 minutes). They were placed in the left lateral position, and the entire esophagus and stomach were visually examined including a careful examination of the lesser curvature of the cardio because it is a frequent site of gastric adenocarcinoma in this region. Next, Lugol's iodine (1.2%) solution was used to stain the full length of the esophagus. The normal esophageal mucosa is brown after iodine staining, while the lesion area is not colored or lightly colored, showing a sharp contrast with the surrounding normal mucosa. The images of typical early esophageal cancer are shown in Figure 1.

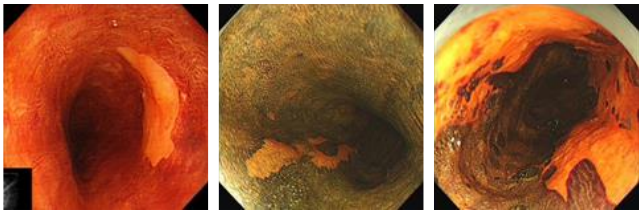


Figure 1. Three typical early esophageal cancer images. The normal digestive tract mucosa is brown after iodine staining, while the lesion area is not stained or lightly colored, and is in sharp contrast with the surrounding normal mucosa.

2.2 Dataset and Preprocessing

The dataset used in this paper comes from 110 patients from April 2016 to May 2018 admitted to the Department of Gastroenterology, the First Affiliated Hospital of Zhengzhou University, Henan Province. The pathological diagnosis has been confirmed by biopsy. The image selection work was assisted by three laboratorians also with the pathological results to confirm the esophageal cancer images. The resolution of the image includes 720×480 pixels, 520×480 pixels, and others. For the sake of saving computing resources, the image size was unified to 500×500 pixels. Besides, the lesion area was marked by two experienced endoscopic physicians and produced to the PASCAL VOC2012 dataset format. The dataset was separated as training set, validation set, and test set. The original dataset has 154 training images, 51 images, 51 test images. The training, validation, and final testing sets of chromoendoscopy images of early esophageal cancer had no overlap. The training set was firstly prepared to train the neural network for the localization and recognition of early esophageal cancer in the TensorFlow, followed by the evaluation of the trained model on test set. Training stopped when the accuracy of validation set is significantly reduced or unchanged. To be robust, a model for recognition and location should make predictions that are unvarying to various inputs. A straight-forward approach is to collect a large number of training samples with abundant variation, regardless of the difficulties in data collection and labeling. Unfortunately, it is always not reality for medical images. Another way to cope with this problem is data augmentation, which is achieved by adding sample replicas with label preservation. According to the types and attributions of the dataset, it was augmented by roughly rotating every 90 degrees and flipping horizontal. The same operations have been done in image labeling process. In addition, since the deep learning network is a high-capacity model, it is useful for avoiding over-fitting.

2.3 Neural Network Architectures

An M-Deeplab model was proposed in this article, with a structure shown in Figure 2. This network was extended from the Deeplabv3+ model by employing an encoder-decoder structure for accuracy improvement. The network consisted of two parts: encoder module and decoder module. Different from the Deeplabv3+ model, the activation function of ReLU has been replaced by SELU for accuracy improvement. The improved Deeplabv3+ model was used as the encoder module to extract the features computed by deep convolutional neural networks, and the atrous convolution controlled the resolution of the feature map under the specified computing resources. The encoder features

were first bilinearly upsampled by a factor of 4 and then concatenated with the corresponding low-level features from the network backbone with the same spatial resolution. The corresponding low-level features usually contain a large number of channels (e.g., 256 or 512) which may outweigh the importance of the rich encoder features (only 256 channels in our model) and make the training harder. Therefore, another 1×1 convolution layer was used in the low-level features for reducing the number of channels. After the connection, a few 3×3 convolutions were applied to refine the features followed by another simple bilinear upsampling by a factor of 4. The output of stride 16 for the encoder module strikes the best trade-off between speed and accuracy. Compared with the Deeplabv3+ model, the improved model had deeper layers, and the maximum pooling layer was replaced by a separable convolution (3×3 with stride 2), and the SELU activation function was added after the 3×3 depthwise convolution to accelerate the model convergence.

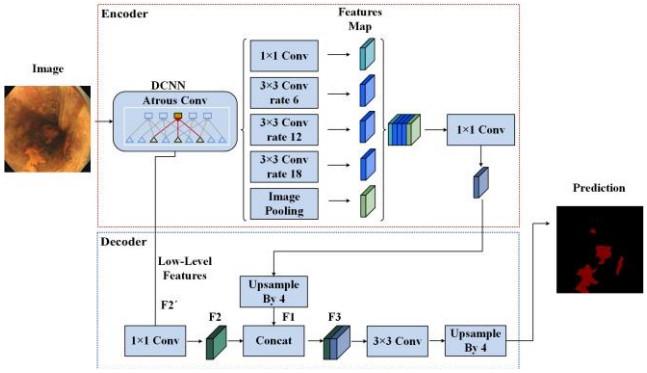


Figure 2. The M-Deeplab model extends Deeplabv3+ by employing an encoder-decoder structure. The encoder module encoded multiscale contextual information by applying atrous convolution at multiple scales, while the simple effective decoder module refines the segmentation results.

2.4 Training

The training process of this model was carried out in the Ubuntu system environment using Python 3.5.2. All experiments were implemented by Tensorflow 1.6.0. software libraries.

The schematic of M-Deeplab training process is shown in Figure 3. The label images were regarded as the ground truth of the lesion, which was marked with the senior doctor's guidance and confirmed by pathological mechanism.

The neural network was randomly specified at the beginning of the training and then was continuously adjusted according to the difference between segmentation map and the actual map until the model converges. During this process, the parameters in the M-deeplab model were optimized to approximate to the ground truth. In this architecture, cross entropy was used as loss function to quantify the error between the output and the actual value. In addition, a large number of fundamental features in the PASCAL VOC2012 dataset were inferred to be transferable to our tasks, and thus were used to pre-initialize the weights before refining the weights in our dataset.

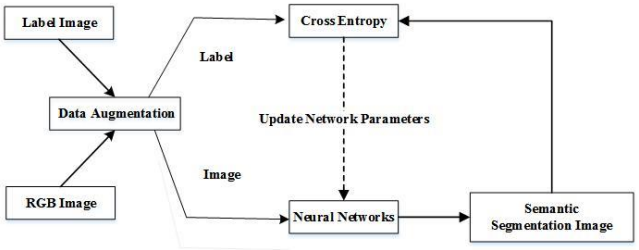
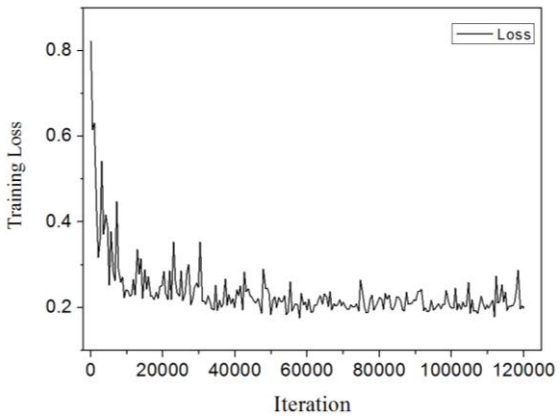
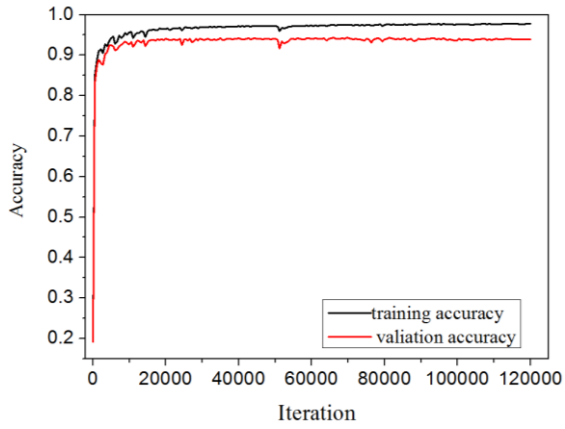


Figure 3. The schematic of training process of M-Deeplab model for early esophageal cancer detection

The Deeplabv3+ model was trained 120k times on the training set and batch_size value was set as 16. The loss function value, training accuracy and validation accuracy of the Deeplabv3+ model in the training phase are shown in Figure 4. When the iteration reached 110k on the training set, the accuracy of the model changes slightly, and then the validation is also prone to saturation.



(a)

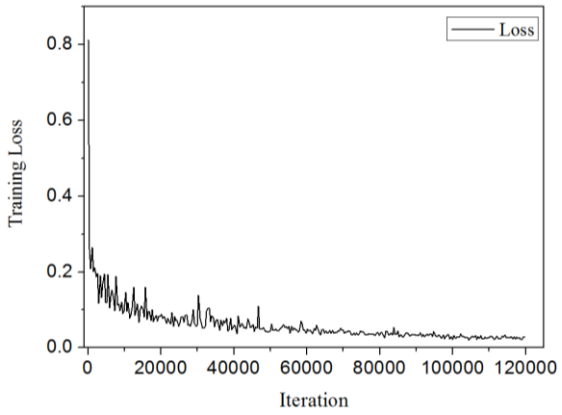


(b)

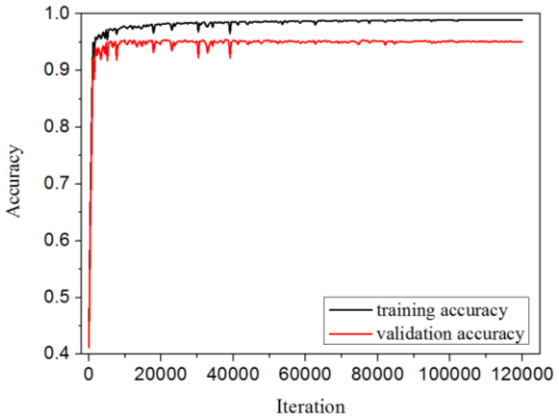
Figure 4. (a) The loss function value and (b) the accuracy of Deeplabv3+ model in the training phase

Based on the Deeplabv3+ network structure, the proposed M-Deeplab structure similarly implemented the training process with an improved activation function of SELU, instead of ReLU in the Deeplabv3+. The loss function value, training accuracy and validation accuracy in the training phase were also shown in Figure 5. When the iteration reaches 100k on the training set, the

accuracy of the model changes slightly, and then the validation is also prone to saturation.



(a)



(b)

Figure 5. (a) The loss function value and (b) the accuracy of M-Deeplab model in training phase

3. RESULT AND DISCUSSION

The Xception model [19] demonstrated fast computational capabilities for image classification on ImageNet [20]. Recently, the Google team improved the Xception model as a Deeplabv3+ model and further promoted the performance of object detection tasks [21]. Driven by these findings, the Deeplabv3+ model will be applied to the task of esophageal cancer segmentation as much as possible. A new model referred to as M-Deeplab has been proposed in this article with the activation function of SELU, substitutes for the function of ReLU in Deeplabv3+ model for accuracy improvement. The experimental results show that the neural network model using SELU activation function has better sparsity than ReLU activation function. The performance index of the model is also improved to some extent.

In order to evaluate how accurate the proposed model, it was compared to other architectures. Different semantic segmentation models were trained and tested, including Fully Convolutional Networks (FCN-32s), Encoder-Decoder with Atrous Separable Convolution (Deeplabv3+), and the M-Deeplab model. The trained model was evaluated on the test set. All the images used in the test were not involved in the training and validation phase, so as to guarantee the test results with unbiased estimations. In the test phase of the algorithm model, the image size of input is the same as the size of the training phase (500×500 pixels). Two key parameters, the accuracy and the Mean Intersection over Union (MIoU), were introduced to effectively evaluate the difference between predicted segmentation and ground truth. They were respectively given by the formula (1) (2).

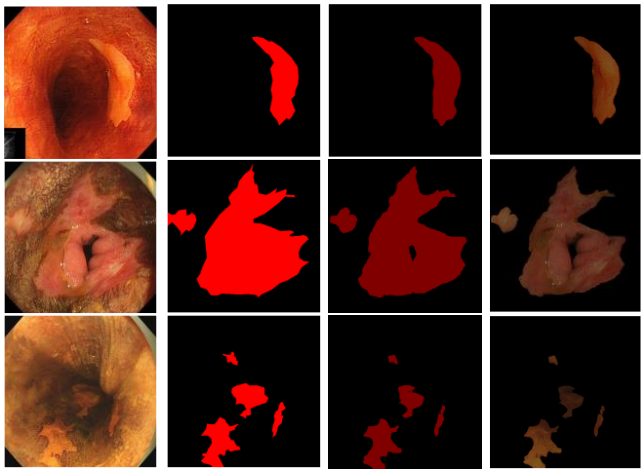
$$Accuracy = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \quad (1)$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \quad (2)$$

Where k is the category, and there are k+1 classes (including the background), P_{ij} indicates the number of pixels that belong to class i but are predicted to be class j. That is to say P_{ii} represents the true quantity, while P_{ij} and P_{ji} are divided into false positive and false negative, both of which are the sum of false positive and negative. The accuracy is the most critical parameter for early esophageal cancer diagnosis, which is defined as the ratio between the correct and the total number on the test set. The MIoU is from the point of view how close of the location region to the ground truth, which is defined as the ratio of intersection between ground truth set and prediction set.

To better evaluate the performance of our model on early esophageal localization and recognition, two other deep learning algorithms of Deeplabv3+, FCN-32s were also assessed at the same dataset in the aspects of accuracy, MIoU and speed. Compared with Deeplabv3+, FCN-32s, M-Deeplab model has the highest overall accuracy of 97.31%. It is also being found that M-Deeplab performs best in the point of view of MIoU, with a remarkable value of 92.09% compared to 90.02% for Deeplabv3+ and 83.76% for FCN-32s. In addition, the M-deeplab takes only 0.05s to judge one image, much faster than the Deeplabv3+ and FCN-32s. Thus, it is convincing that the performance of M-Deeplab model outperforms the other methods on our dataset both in accuracy, MIoU and speed. By further study, the M-Deeplab model has been proved to reach the similar accuracy level with human experts. These results show that scaled exponential linear units (SELU) can make a positive contribution for semantic esophageal cancer image segmentation.

The test results of M-deeplab model on 100k iterations are shown in Figure 6. The first column represents the test image of input; the second column shows the ground truth image that is labeled by experienced experts and proved by pathological mechanism; the third column is the prediction image of the M-Deeplab model; and the fourth column indicates the prediction lesions on the original image. It can be seen that the M-Deeplab model predicts accurately in the early esophageal cancer recognition, and locates precisely for the pathological changes region. This can provide effective auxiliary diagnostic information for early esophageal cancer diagnosis and at the same time help doctor to reduce the work intensity.



(a) image (b) ground truth (c) M-Deeplab (d) M-Deeplab-masked

Figure 6. Semantic segmentation results of lesion area based on M-Deeplab model

The deep-learning method proposed in this paper effectively solves the problem of quantitative analysis of fixed rectangular regions, and plays an important role in future biopsy and other pathological examinations. At present, the algorithm can only run on the remote server. The device terminal uploads the collected image data to the server and then returns to the segmentation result. Therefore, in the future research, the algorithm will be deployed directly to the device terminal. In addition, further research is also expected to increase the number of images used in the study.

Table 1. Overall accuracy, MIoU, and speed of main component in this proposed method

Model	Overall accuracy	MIoU	image speed
M-Deeplab	97.31%	92.09%	20 frames/s
Deeplabv3+	95.77%	90.02%	17 frames/s
FCN-32s	94.09%	83.76%	1 frames/s

4. CONCLUSION

In this study, an M-Deeplab model was proposed for highly accurate locating and recognizing early esophageal cancer. Experiments show that the M-Deeplab model can achieve 97.31% pixels accuracy and 92.09% MIoU in the test dataset. The proposed method also has better performance than original Deeplabv3+ and FCN models. It can be concluded that the deep learning algorithm can enhance the sensitivity of lesion identification, reducing the errors in subjective judgment at the same time. Our model can serve as a clinical assistance for guiding the location of biopsy tissue, and improving the diagnosis rate of early esophageal cancer.

5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61874099 and No.81802325).

6. REFERENCES

[1] Whiteman, D. C. 2014. Esophageal cancer: priorities for prevention. *Current Epidemiology Reports, 1*(3), 138-148.

[2] Kandiah, K., Chedgy, F. J., Subramaniam, S., Thayalasekaran, S., Kurup, A., & Bhandari, P. 2017. Early squamous neoplasia of the esophagus: the endoscopic approach to diagnosis and management: Saudi Journal of Gastroenterology Official Journal of the Saudi Gastroenterology Association, 23(2), 75-81.

[3] Uedo, N., Fujishiro, M., Goda, K., Hirasawa, D., Kawahara, Y., & Lee, J. H., et al. 2011. Role of narrow band imaging for diagnosis of early‐stage esophagogastric cancer: current consensus of experienced endoscopists in asia–pacific region. Dig Endosc, 23(s1), 58-71.

[4] Ajani JA, Barthel JS, Bentrem DJ, D'Amico TA, Das P, & Denlinger CS, et al. 2015. Esophageal and esophagogastric junction cancers. J Natl Compr Canc Netw, 13(2), 194-227.

[5] di Pietro M, Canto MI, Fitzgerald RC. Endoscopic Management of Early Adenocarcinoma and Squamous Cell Carcinoma of the Esophagus: Screening, Diagnosis, and Therapy. Gastroenterology. 2018 Jan;154(2) 421-436.

[6] Mannath, J., & Ragunath, K. 2016. Role of endoscopy in early oesophageal cancer. *Nat Rev Gastroenterol Hepatol, 13*(12), 720-730.

[7] Roth, M. J., Liu, S. F., Dawsey, S. M., Zhou, B., Copeland, C., & Wang, G. Q., et al. 2015. Cytologic detection of esophageal squamous cell carcinoma and precursor lesions using balloon and sponge samplers in asymptomatic adults in linxian, china. *Cancer, 80*(11), 2047-2059.

[8] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., & Narayanaswamy, A., et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama, 316*(22), 2402.

[9] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, & Helen M. Blau, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118.

[10] Na Niu, WJ Zhang, XF Lu, et al.2012. Multicolor spatial quantitative analysis for the diagnosis of early esophageal cancer [J]. Television Technology, 36 (11): 134-137.

[11] Wu Ye, PF Liu, XM Cao, et al. 2014. Value of computer-assisted chromoendoscopy in the diagnosis of early esophageal cancer [J]. World Chinese Journal of Digestology, 3811-3814.

[12] Schmidhuber, J. 2015. Deep learning in neural networks: an overview. *Neural Netw, 61*, 85-117.

[13] Baldi, P., & Chauvin, Y. 1993. *Neural networks for fingerprint recognition*. MIT Press.

[14] Graves, A., Mohamed, A. R., & Hinton, G. 2013. Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol.1, pp.6645-6649). IEEE.

[15] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., & Macherey, W., et al. 2016. Google's neural machine translation system: bridging the gap between human and machine translation.

[16] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., & Van, d. D. G., et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature, 529*(7587), 484-489.

[17] Di, L. P., Nagata, K., & Baldi, P. 2012. Deep architectures for protein contact map prediction. *Bioinformatics, 28*(19), 2449-2457.

[18] Wei, W. Q., Chen, Z. F., He, Y. T., Feng, H., Hou, J., & Lin, D. M., et al. 2015. Long-term follow-up of a community assignment, one-time endoscopic screening study of esophageal cancer in china. *Journal of Clinical Oncology, 33*(17), 1951-1957

[19] Chollet, F. 2016. Xception: deep learning with depthwise separable convolutions. 1800-1807.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, & Sean Ma, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211-252.

[21] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. http://arxiv.org/abs/1802.02611

Automatic Recognition for Arbitrarily Tilted License Plate

Nanxue Lu
University of Science and Technology of China.
lnx1212@mail.ustc.edu.cn

Wei Yang
University of Science and Technology of China
qubit@us tc.edu.cn

Ajin Meng
University of Science and Technology of China
sa516210@mail.ustc.edu.cn

Zhenbo Xu
University of Science and Technology of China.
xuzhenbo@mail.ustc.edu.cn

Huan Huang
Xingtai Financial Holdings Group Co., Ltd.
huanghuan@xtkg.com

Liusheng Huang
University of Science and Technology of China.
lshuang@ustc.edu.cn

ABSTRACT

In this paper, we propose a novel automatic license plate recognition (ALPR) method based on convolutional neural network to achieve a better performance in detecting and recognizing license plate (LP) with relatively large angle of inclination. Most existing methods only perform well on dataset where LPs are presented in almost upright position with little or no tilted angle. While, in practice, the LP images collected by roadside cameras or hand-held image capturing devices can be fairly slanted, which causes great difficulties on recognition tasks. To solve this problem, we design an angle correction module and integrate it into a holistic ALPR model with a spatial transformer network embedded inside. The whole model can be trained end-to-end by back-propagation. A large and comprehensive rotated LP dataset Rlpd is collected and introduced in our work for model training and testing. Through extensive experiments, this approach is proved to have a better performance on tilted license plate dataset in terms of accuracy and computational cost than other state-of-the-art methods.

CCS Concepts

• Computing methodologies → Neural networks.

Keywords

Car License Plate Detection and Recognition, Intelligent Transportation System, Computer Vision, Convolutional Neural Network.

1. INTRODUCTION

number of private cars on the road is surging, transportation systems become more and more vulnerable to traffic problems. For a favorable traffic environment, we need to keep all vehicles under careful surveillance, which requires automatic license plate recognition (ALPR) technology to help identifying vehicles.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301547

However, the LP images collected by roadside cameras or hand-held devices are of poor quality for most of the time. The LP in an image can be tilted, blurred or under poor illumination condition. We cannot correct all of these distortions manually because the volume of data is huge and manpower is costly. So we need a robust ALPR approach to handle LP images with transformation.

Our work in this paper mainly focuses on recognizing LPs oriented in various directions. Most exiting ALPR methods don't give enough attention to this problem. In practice, even the most popular ALPR techniques based on convolutional neural network (CNN) still suffer from low recognition accuracy when encountering largely tilted LPs due to CNN's lack of invariance to transformation. Although spatial transformer network [9] is proposed to address this issue, problems still exist. We find it hard to make the network embedded with spatial transformer converge when training on our proposed rotated LP dataset Rlpd.

In order to improve the ALPR performance on tilted LP dataset, we introduce a novel ALPR approach combining CNN with a detector and classifier module (DCM), an angle corrector module (ACM) and a spatial transformer module (STM) [9] to handle slanted LPs. DCM regresses the bounding box (bbox) of the LP and predicts the probable rotated angle of a LP. Then ACM roughly corrects the inclination. STM is inserted after ACM to further adjust the tilted angle and accelerate the training process. The model is trained on Rlpd. In experiments, our method shows great capability in handling largely oriented LPs and is proved to reach high recognition accuracy close to that of the human eye.

Both Rlpd and the code for training and evaluating our model are available under the open source MIT license at: https://github.com/ALPRRlpd/Rlpd.

The main contributions in this paper are summarized as follows:

- We design a novel CNN-based architecture combining an ACM and a STM to accomplish LP detection and recognition tasks holistically on LPs with large angle of inclination. This model has strong robustness to LP rotation and can be trained end-to-end using back-propagation.
- By try and error, we explore the advantages and limitations of employing a spatial transformer in an ALPR model to increase the recognition accuracy and the converging speed. It is proved that selecting an appropriate input size of the spatial transformer improves the performance of the whole model to a large extent.

- We introduce Rlpd, a large and comprehensive rotated license plate dataset publicly available for ALPR to date. Through extensive experiments, we demonstrate that our method combining STM and ACM has better performance than other methods and can achieve state-of-the-art recognition accuracy on slanted LP dataset.

2. RELATED WORK

LP recognition under various circumstances has been studied over the last few decades; comprehensive surveys can be found in [1][4]. In this chapter, we briefly review some related literatures.

Traditional ALPR methods adopt classical hand-crafted features to help locating and identifying LPs [2][15]. It is difficult for them to handle transformation because the features they rely on are too simple and vulnerable to distortion. We need more abstract features from higher dimensions that are more invariant to transformation. Convolutional Neural Network (CNN) is introduced to ALPR techniques, which is suitable for processing grid data like images. [7] and [5] use CNN for image semantic segmentation. AlexNet [10], the champion of 2012 ImageNet competition, introduces CNN into image classification task. The ALPR methods based on CNN has reached a better performance than many other traditional approaches. [11][3] train a deep CNN classifier to extract sequential features from LP images and reach state-of-the-art LP recognition accuracy. Although CNN is powerful, it is still not totally invariant to transformation including translation and rotation [9]. When the input LP images are largely distorted by the deviation of cameras, CNN-based models fail to correctly detect and recognize LPs lie inside. For arbitrarily tilted LPs, existing method can detect it from an image [14], but there is few research focusing on recognizing this kind of LPs. To the best of our knowledge, so far, no effective method is proposed specifically for recognizing tilted LPs especially when the angle of inclination is large. The approach in [13] performs well on regular dataset but the author points out that it has little tolerance for LP rotation (less than 30 degree).

Most traditional ALPR methods are comprised of three steps: LP detection, characters segmentation and characters recognition [4]. This type of ALPR approaches has a notable drawback. Its performance largely relies on the robustness of each individual stage. [13] proposes a segmentation-free ALPR method based on CNN. It uses several CNN layers to extract features of LP first and then feeds them into fully-connected (FC) layers with seven branches. Each branch serves as the specialized classifier for a character on a specific position of the input LP image. This method avoids the possible errors generated in segmentation step and supports end-to-end training. Inspired by this "holistic" idea, our method is also designed as an indivisible process which is easy to optimize.

3. FRAMEWORK

In this paper, we propose a novel CNN-based ALPR method specialized in recognizing arbitrarily-oriented LPs. The architecture we design consists of four modules: detector and classifier module (DCM), angle corrector module (ACM), spatial transformer module (STM) and recognizer module (RM) as shown in Figure 1. DCM is a pre-trained neural network with two outputs, bbox regression and orientation classification, aiming at extracting LP from the input image and classify it into four types

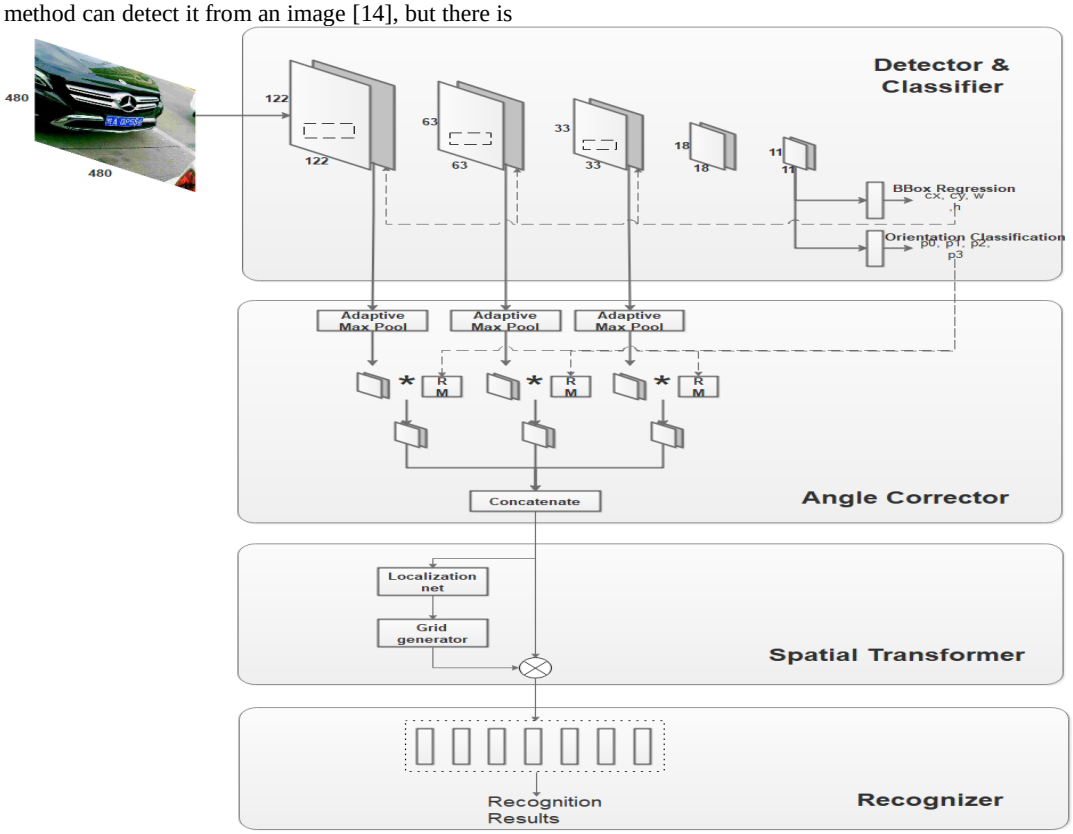


Figure 1. The framework of our model.

by its inclined angle (Section 4.1). ACM receives the result from DCM and corrects the tilted LP based on its feature maps catenation and the orientation type categorized by DCM (Section 4.2). STM further processes the output from ACM and accelerates the converging speed (Section 4.3). RM takes the corrected feature map from last step and outputs the final result of identification (Section 4.4).

Firstly, the input RGB image is passed to DCM to extract convolutional features, regress bbox parameters and classify the extracted LP into four classes by its rotated angle. Secondly, low-level feature maps and angle type prediction are delivered to ACM. ACM resizes the feature maps into a fixed size (16×16 in our work) using adaptive max pooling [6]. Then, we perform rotation correction on feature maps by multiplying special matrices to adjust their tilted angle. The result is fed into STM where the feature maps are further corrected. Lastly, RM identifies the LP by classifying characters appeared on each position into possible categories and output the final result.

4. METHODOLOGY

4.1 Detector and Classifier Module (DCM)

Although detecting a LP from an image is not part of this work, we mention it here for completeness of the chapter. DCM is a pre-trained network with ten convolutional layers in a sequence and two parallel fully-connected layers. The input image of fixed size 480×480 goes through eight convolutional layers with 5×5 filters and two convolutional layers with 3×3 filters. The output is then fed to two FC layers, bbox regression layer and orientation classification layer.

4.1.1 BBox Regression Layer (BRL)

An input image always contains information other than LP. So first of all we locate the LP by inferring the position and size of its bbox. BRL produces four values as the bbox prediction of the LP (cx, cy, w, h) where cx, cy represent the coordinate of the left-top point of the bbox and w, h are the width and height of the bbox.

4.1.2 Orientation Classification Layer (OCL)

A LP in an image can rotate by any angle. Here we only classify LP into four categories 0, 90, 180, 270 according to the angle they rotate. In other word, we don't consider about the exact angle a LP is tilting but roughly decide the group they belong to. Figure 2 shows an example of how we define the rotation type of a LP. From left to right, the classified angle types are 0, 90, 180, 270 degrees clockwise. OCL produces four values representing the possibility of the classification result of LP's tilted angle.

BRL and OCL share the same convolutional layers to extract features from images, which enables us to save storage space of the entire model and accelerate the training process.

Figure 2. Examples of four types of orientation.

4.2 Angle Corrector Module (ACM)

Region of Interest (RoI) is the valid region that we focus on in an input image. [6] proposes an RoI pooling layer to convert the features inside any RoI into a small feature map. ACM is the combination of RoI pooling layer and a matrix rotation strategy. It receives three low-level feature maps from the second, fourth and sixth convolutional layer from DCM as well as the bbox regression and rotation type prediction. Then, RoI on each feature map is cropped out and resized to the same size (16 × 16 in this paper) by adaptive max pooling. According to the prediction of tilted angle, corresponding rotation operations are performed on each feature map to roughly correct the tilted angle. For instances, in equation (1), suppose matrix M is a feature map extracted from a LP image in which the LP tilts by around 270 degrees clockwise and RM is a rotation matrix. $M^T * RM$ corrects M in the way of rotating it by 90 degrees clockwise.

$$M = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \quad RM = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$M^T * RM = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} * \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} g & d & a \\ h & e & b \\ i & f & c \end{bmatrix} \quad (1)$$

4.3 Spatial Transformer Module (STM)

The feature maps produced by ACM are already corrected to some extent in terms of the inclined angle of the LP. And by experiments, we can see a significant increase on recognition accuracy. However, in real world, a LP presented in an image can rotate in any angle, and here in our work, we just roughly classify the observed angles into four categories by the closeness between value of the actual angle and the four "criterion" angles, 0, 90, 180, 270 degrees. All the other angles besides the four "criterion" angles are forced to belong to one of the four classes. Therefore, after doing angle correcting step, it is possible that the LP still has slight rotation angle. In practice, the slight rotation angle still throws a negative impact on LP recognition accuracy.

To further improve the performance of our model, we embed a spatial transformer module (STM) [9] after ACM. As shown in Figure 2, a STM is the combination of a localization network, a grid generator and a sampler. This module helps the network to learn how to actively transform the feature maps to an appropriate state so that the overall cost of the network can be minimized during training. This knowledge is compressed and cached in the weights of the localization network and the layers previous to a spatial transformer during training.

In our model, STM helps to further adjust the tilted angle of a LP after the rough correction done by ACM. Theoretically, STM itself can accomplish the inclination correction task without cooperating with ACM, but when the rotated angle of a LP is too large, it is very hard for the parameters of the whole network to converge and the training speed is slow. Through experiments, we find out the combination of ACM and STM performs better than applying them separately, which is described in detail in Section 5.

4.4 Recognizer Module (RM)

Most traditional ALPR methods perform LP extraction, character segmentation and LP recognition in separate steps. The performance of one step is easily affected by the output of previous steps. This separation also makes improvement on the total performance of the entire model very difficult.

RM takes full advantage of the holistic ALPR approach [13]. First, except for the first Chinese character, characters on a LP can be divided into 37 types, 26 uppercase English letters, 10 digits and a negative non-character category. Since we are coping with Chinese LPs with the length of 7 characters, the network in RM has 7 branches and six of them have 37 types of outputs predicting

one character on one of the six positions on a LP. The branch for predicting the first Chinese character 34 types of outputs each of which standing for a province in China. In this manner, we don't need to separate the recognition process into several steps, which makes optimization much easier than other separated approaches.

4.5 Loss Functions

As we demonstrate previously, the network takes an RGB image, its bbox and ground truth label as the input during training time and realizes detection and recognition in one shot. So we employ the multi-task loss to jointly tune the model parameters. The loss of the whole model consists of three parts, the loss of bbox regression L_{bbox} ,the loss of LP orientation type classification L_{orien} and the loss of LP character prediction L_{plate} :

$$L_{bbox} = L1((cx, cy, w, h), (cx^{gt}, cy^{gt}, w^{gt}, h^{gt})) \quad (2)$$

$$L_{orien} = CrossEntropy((o_0, o_1, o_2, o_3), orien^{gt}) \quad (3)$$

$$L_{plate} = \sum_{i=0}^{6} CrossEntropy((p_0, p_1, ..., p_{c_i}), p_i^{gt}) \quad (4)$$

So the multi-task loss function is defined as the addition of all three losses:

$$L = \frac{1}{N}(L_{bbox} + L_{orien} + L_{plate}) \quad (5)$$

where N is the volume of a training batch. In equation (2), the definition of bbox loss is the L1-norm distance between the predicted bbox (cx, cy, w, h) and the ground truth $(cx^{gt}, cy^{gt}, w^{gt}, h^{gt})$. In equation (3), cross-entropy loss is used to represent orientation classification loss, where o_0 , o_1 , o_2 , o_3 denote the probability of the tilted angle of an LP being assigned to one of the four classes. The loss of plate characters (4) is also the cross-entropy loss, where p_0 , p_1 ,..., p_{c_i} are probabilities of each character on an LP being assigned to the 37 classes.

5. EXPERIMENTS

In this section, we conduct experiments to verify the effectiveness of our proposed methods. Our network is implemented using Pytorch 0.4. The experiments are performed on eight 3.40 GHz Intel Core i7-6700 CPUs, 24GB RAM and one Quadro P4000 GPU. The code and Rlpd dataset are available on github.

5.1 Dataset Acquisition

5.1.1 Rlpd Generation

We collect a new dataset called Rlpd (Rotated license plate dataset) where LP images are recorded by some toll collectors using cellphone's camera under different conditions. LPs in Rlpd are all Chinese LPs as shown in Figure 3. From left to right, the first is a Chinese character, then an English letter followed by a sequence of digits or English letters. The first Chinese character has 34 types representing 34 different regions in China. In our dataset, years of LP recordings come from around 400 different locations in various daytimes. We randomly crop the images to further augment the dataset. Then some LP images of small rotation angle (less than 90 degrees) are chosen and rotated by 90, 180 or 270 degrees to get large-degree tilted images. The volume of Rlpd is 400K, which we sample into three parts, 380K for training, 10K for evaluating and the rest 10K for testing.

Figure 3. Examples of some Chinese license plates.

5.1.2 Ground Truth Assignment

The collected LPs are labeled by their ground truth texts comprised of 3 parts, LP number on each LP assigned and uploaded to the database by toll collectors, coordinate of bbox annotated by a web-based annotation tool and rotation angle of a LP recorded when generating tilted LPs from non-rotated LPs.

5.2 Training

The training process can be divided into two phases: Pre-train DCM. The LP detector and orientation angle classifier need to be trained first because the bbox regression and orientation type prediction produced by this part greatly affect the performance of the whole model. Also, training DCM separately can largely accelerate the converging process; Train the whole network. We use transfer learning to train the whole model. The first few convolutional layers are initialized from the pre-trained model and the rest of the parameters are initialized in a random manner. Then, Stochastic Gradient Descent is applied to minimize the loss.

5.3 Comparative Analysis

In this part, we construct four models, holistic ALPR model (H-ALPR), holistic ALPR with DCM and ACM (H-ALPR-DC-AC), holistic ALPR with STM (H-ALPR-ST) and holistic ALPR with DCM. ACM and STM (H-ALPR-DC-AC-ST). These four models are all fine-tuned, evaluated and tested on Rlpd dataset.

After training, the recognition accuracy of the four models is shown in Table 1 where we use two criteria to evaluate these models, all correct accuracy (ACA) and six correct accuracy (SCA). As is stated before, A Chinese LP contains 7 characters including a Chinese character, an uppercase English letter and a combination of five characters including digits and uppercase English letters. Chinese characters only appear on Chinese LPs but English letters and digits can be seen on LPs of many other countries. ACA describes the accuracy of correctly recognizing all the seven characters on a LP while SCA describes the accuracy of correctly recognizing six characters except for the Chinese character. These two recognition tasks vary in the degree of difficulty. Generally speaking, recognition for Chinese characters is more challenging.

Table 1. Recognition accuracy of the four models.

	ACA(%)	**SCA(%)**
H-ALPR	89.24	96.78
H-ALPR-DC-AC	91.14	97.52
H-ALPR-ST	0	0
H-ALPR-DC-AC-ST	**94.47**	**98.18**

Figure 4 shows the loss function curve of the four models during training, from which we can see H-ALPR-DC-AC-ST converges much faster than another three models while the loss of H-ALPR-ST stops going down after the early few rounds of training. Figure 5 shows the accuracy curve of the four models during training. The accuracy of model H-ALPR-DC-AC-ST rises much faster than another three models and finally reaches the highest point. And the accuracy of model H-ALPR-ST remain close to zero because it doesn't converge.

5.3.1 H-ALPR

This model doesn't contain any module for LP inclination correction. The feature maps output from the second, fourth and sixth convolutional layer go directly into RM after being resized and concatenated. It can only tolerate the LP tilting in a very small angle {less than 30 degree} [13]. When we apply it on Rlpd, its recognition accuracy decreases a lot especially for identifying the first Chinese character on a LP.

5.3.2 H-ALPR-DC-AC

Compared with H-ALPR, H-ALPR-DC-AC makes some improvement. This model integrates DCM and ACM into H-ALPR. DCM regresses the bbox of LP and gives prediction on its slanted angle type. Then ACM roughly corrects the inclined angle of an LP by rotating the it against its tilted direction by 0, 90, 180, or 270 degrees, which makes the LP easier to identify and therefore, increases the recognition accuracy.

5.3.3 H-ALPR-ST

This model combines H-ALPR with STM expected to learn transformation parameters automatically and minimize the total loss of the model. Theoretically, it can handle any kind of distortion of target object including large-angle rotation of the LPs in ALPR system. But in experiment, the loss stops decreasing after a few epochs of training as shown in Figure 4. When the target object is distorted a lot, it is difficult to make the model converge using only STM to adjust the inclination.

5.3.4 H-ALPR-DC-AC-ST

This is the model we propose in our work applies DCM, ACM and STM together in H-ALPR. DCM performs LP detection, bbox regression and tilted angle classification. ACM then roughly corrects the inclination using information receives from DCM. After that, STM further fine-tunes the tilted angle of the LP. In this case, recognition accuracy increases and the converging speed is accelerated significantly as shown in Figure 4.

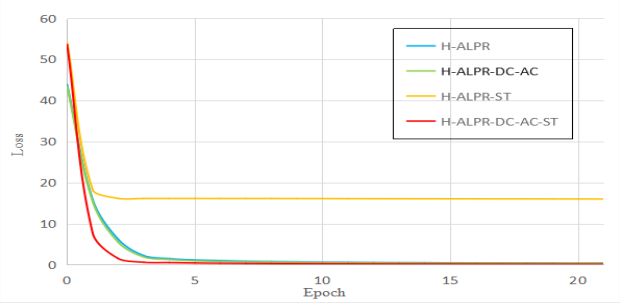


Figure 4. Loss function curve of the four models.

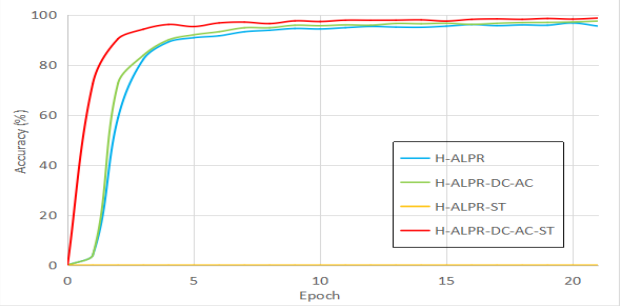


Figure 5. Accuracy curve of the four models.

As it is illustrated in [9], the input size of STM is fixed. We try 16 × 16, 32 × 32 and 64 × 64 three types of size to find the most appropriate one. For 16 × 16, the features captured by the network is not enough for recognition and the accuracy decreases. 64 × 64 feature map maintains most features delivered from previous layers, but respectively, the output size of STM grows and causes the number of parameters in the following FC layers increase exponentially, which makes the whole network difficult to converge. It turns out that 32 × 32 input performs best both in recognition accuracy and converging performance.

6. CONCLUSION

This paper results in a CNN-based ALPR that can detect and recognize LPs with large angle of rotation. The system is tested on real-life Chinese LP image dataset collected by roadside toll collectors as Rlpd. The LP is recognized correctly in 95% images, which is superior to that of human eyes and close to the state-of-the-art. The proposed combination of ACM and STM forms an accurate LP recognition system robust to LP inclination. The employment of transfer learning largely reduces the training time since part of the layers of the networks are already trained. When performing training and testing on a GPU, the overall system processes 40 images per second. As is usual for deep learning based computer vision systems, our system is specially adjusted to work on GPU. Our future work involves further improvement of the LP recognition accuracy on different kinds of transformation and optimization of the system for mobile devices.

7. REFERENCES

[1] C. E. Anagnostopoulos, I. E. Anagnostopoulos, I. D. Psoroulas, V. Loumos and E. Kayafas, "License Plate Recognition From Still Images and Video Sequences: A Survey," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 3, pp. 377-391, Sept. 2008.

[2] Shyang-Lih Chang, Li-Shien Chen, Yun-Chung Chung, and Sei-Wan Chen. 2004. Automatic license plate recognition. *Trans. Intell. Transport. Sys.* 5, 1 (March 2004), 42-53.

[3] Nauris Dorbe, Aigars Jaundalders, Roberts Kadikis, and Krisjanis Nesenbergs. Fcn and lstm based computer vision system for recognition of vehicle type, license plate number, and registration country.

[4] Shan Du, Mahmoud Ibrahim, Mohamed Shehata, and Wael Badawy. Automatic license plate recognition (alpr): A state-of-the-art review.

[5] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Scene parsing with multiscale feature learning, purity trees, and optimal covers. CoRR,, abs/1202.2160,2012.

[6] Ross Girshick. Fast r-cnn. arXiv preprint arXiv:1504.08083, 2015.

[7] David Grangier, Lon Bottou, and Ronan Collobert. Deep convolutional networks for scene parsing. abs/1411.4101

[8] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *European Conference on Computer Vision*, 2014.

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu. Spatial transformer networks. CoRR, 2015.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors,

[11] Hui Li, Peng Wang, Mingyu You, Chunhua Shen. Reading car license plates using deep neural networks. 72, 03 2018.

[12] David G Lowe. Object recognition from local scale-invariant features. In Computer vision, 1999. The *proceedings of the seventh IEEE international conference* on, volume 2, pages 1150 – 1157. Ieee, 1999.

[13] J. Špaňhel, J. Sochor, R. Juránek, A. Herout, L. Maršík and P. Zemčík, "Holistic recognition of low quality license plates by CNN using track annotated data,"

[14] L. Xie, T. Ahmad, L. Jin, Y. Liu, and S. Zhang. A new cnn-based method for multi-directional car license plate detection.

[15] Khalil M Ahmad Yousef, Maha Al-Tabanjah, Esraa Hudaib, and Maymona Ikrai. Sift based automatic number plate recognition.

Real-Time Face Attendance Marking System in Non-cooperative Environments

Kai Jin
School of Artificial Intelligence
Xidian University
Xi'an, Shaanxi, China
jink.xidian@gmail.com

Xuemei Xie
School of Artificial Intelligence
Xidian University
Xi'an, Shaanxi, China
xmxie@mail.xidian.edu.cn

Fangyu Wang
School of Artificial Intelligence
Xidian University
Xi'an, Shaanxi, China
wfy199508@gmail.com

Xu Gao
School of Artificial Intelligence
Xidian University
Xi'an, Shaanxi, China
xdxuge@163.com

Guangming Shi
School of Artificial Intelligence
Xidian University
Xi'an, Shaanxi, China
gmshi@xidian.edu.cn

ABSTRACT
Face recognition achieves good performance in attendance marking system, but most of face attendance marking systems need people cooperate with the camera. In this paper, we propose the real-time face attendance marking system. It works well in non-cooperative environments. Firstly, detected face regions are tracked to help detection algorithm detect occluded and deformed faces. Then, the features of tracklets (track fragment) formed by detected faces are extracted to realize face recognition. By using tracklets instead of a single image, it realizes robust recognition in non-cooperative environments. Finally, we create a reference gallery of multimodal facial features, which improves the accuracy and speed of multimodal face recognition. Experiments show that our system can detect and recognize multimodal faces in non-cooperative environments and run in real-time (25FPS).

CCS Concepts
• **Information systems → Data analytics**

Keywords
Attendance marking system, Real-time, Cooperative environments, Face recognition, Reference gallery

1. INTRODUCTION
With the development of deep learning, impressive performance has been achieved on face recognition[5][7][11][13][14]. Face recognition has been applied to attendance marking system and achieved good performance. However, most of these systems need people cooperate with camera. It is difficult for these systems to recognize faces that are occluded, blurred or deformed. So, how to design a real-time face attendance marking system used in non-cooperative environments is a challenge.

Several methods have been proposed to build attendance marking systems[2][4][6][8][9]. In these systems, face detection and face recognition are unified to realize face attendance marking. The methods of these systems perform well in cooperative environments, but not well in non-cooperative environments. As for face detection, some occluded or deformed faces are difficult to be located, which causes missing detection. As for face recognition, it is hard to precisely recognize the faces that are heavily deformed. We propose a method that unifies face detection, face tracking and face recognition to realize face attendance marking in non-cooperative environments. In this paper, we adapt optimized MTCNN[15] to detect faces. The detected faces are tracked to form tracklets, in order to help MTCNN detect deformed and occluded faces. Moreover, we extract features of tracklets and take the category with the highest confidence as the face label.

As for the face reference gallery, the method of the establishment affects the accuracy and speed of face recognition in a real scenario. Some reference galleries[2][4][8] contain one face image per person, while others contain many face images per person[9]. For the former, although the system using these galleries have faster speed in recognition, recognition precision is limited as heavily deformed images are largely different from the images in gallery. For the latter, the systems using those galleries perform better on deformed faces, but the reference galleries have too many redundant face images which leads to poor efficiency. In order to realize the accurate deformed face recognition and real-time performance, we establish the multimodal face feature reference gallery. The gallery contains five face images' features from five different angles per person. Meanwhile, our reference gallery stores image features extracted by CNN instead of images. So, the features don't need to be extracted repeatedly for recognition and a lot of time will be saved.

This paper is organized as follows. Section 2 briefly introduces the related work. The algorithm of the system is discussed from Section 3.1 to 3.3. The establishment of face feature reference gallery is

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong.

ACM ISBN 978-1-4503-6613-7/18/12…$15.00.
https://doi.org/10.1145/3301506.3301546

described in Section 3.4. Experimental results are shown in Section 4. Section 5 draws a conclusion of the paper.

2. RELATED WORK

2.1 Face Attendance Marking Method

The existing face attendance marking methods mainly consist of face detection and classification algorithm. Chintalapati et al.[2] proposes to use Viola-Jones detection algorithm for face detection, then applies local binary pattern histogram (LBPH) for feature extraction and support vector machine(SVM) for classification. Sajid et al.[6] proposes a model uses an integral validation process, which enhances the reliability of facial recognition. Selvi et al.[8] proposes to utilize illumination invariant algorithm to remove the lighting effect of the environment. Sharma et al[9] tries to apply error correcting output codes(EOOC) in classification to enhance the precision of the classification system. In our method, we add tracking algorithm between detection and recognition algorithm, which tracks the face when it is detected and recognizes the face multiple times during the tracking process. Then we take the category with the highest confidence as the face label.

2.2 Establishing Reference Gallery

There are two main methods to establish face feature reference gallery. The camera captures a single ideal face image or a large number of face images of different angles as a database. In [2] and [8], The images are captured and pre-processed by Histogram Normalization technique to establish the face reference gallery. Sharma et al.[9] captures 200 images of different angles for each person and stores them as face gallery. In our method, we first collect face images of five different angles through the camera, and then store the extracted features to establish our face gallery, which ensures the accuracy of face recognition and improves the recognition efficiency.

3. PROPOSED SYSTEM

As shown in Figure 1, the system consists of four parts: video capturing, face detection, face tracking and face recognition. The input of the system is the real-time stream from the camera. Our camera is placed at the building entrance, which can recognize the people entering the building and realize attendance marking.

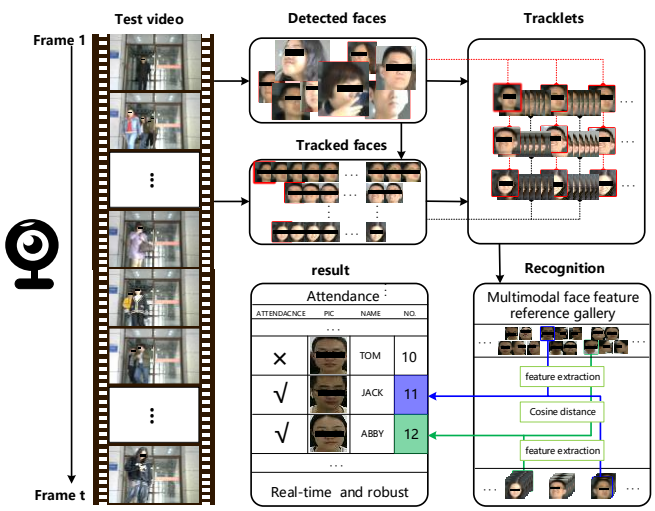


Figure 1. Overview of the proposed system. We hide eyes to protect the privacy of people.

In this system, firstly, frames are captured in the video stream. Then we input frames into the optimized MTCNN to detect faces, and tracking is used to help detect faces that are occluded and deformed. After that, tracklets formed by detected faces are input into face recognition network and the face features extracted by recognition network. Finally, face features are compared with those in face feature reference gallery, and the category with the highest confidence is selected as the face label.

3.1 Face Detection

In our face attendance system environment, the faces are mainly distributed in large and medium size. Original MTCNN[15] is used to detect not only large and medium faces, but also small faces. It causes objects that are not faces to be recognized as small faces in our environment, which causes high false negative rate. Besides, appropriate face confidence thresholds have an impact on detecting large and middle faces in our environment. So, we apply the optimized MTCNN to detect faces, which is a deep multi-task

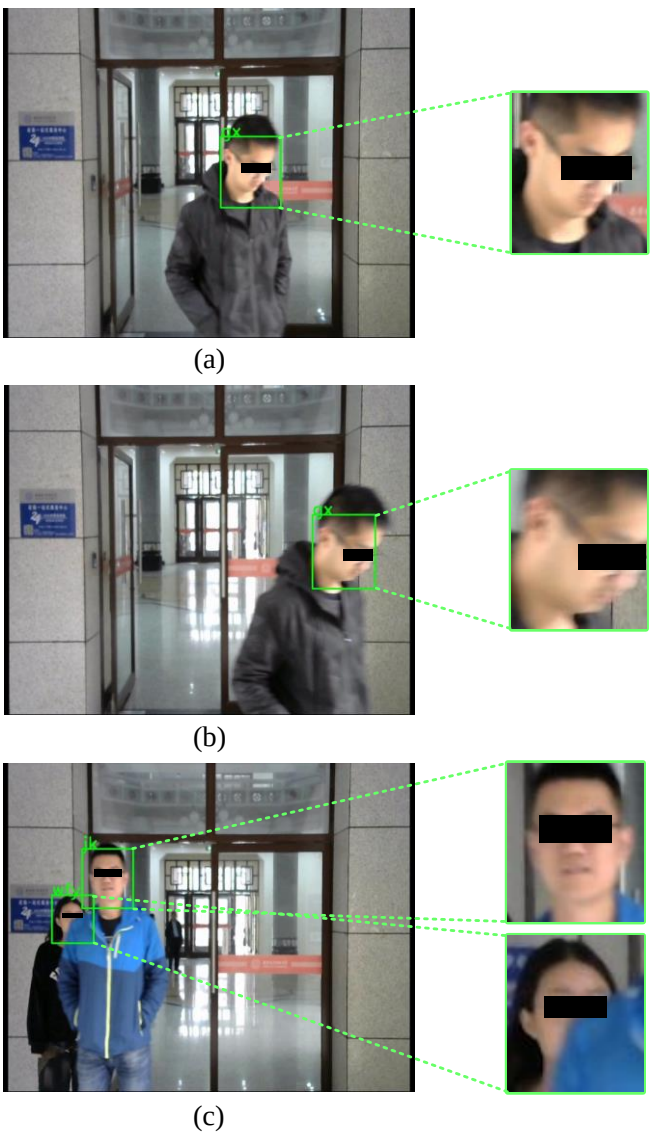


Figure 2. The result of detection with tracking in different condition (a) deformable faces (b) motion blur (c) occlusion.

framework cascaded by P-Net, R-Net and O-Net. It predicts face label and landmark location in a coarse-to-fine manner. As the tiny face is useless in our case, we set the minimum side length to 50 pixels and discard the detected faces smaller than it. Then, we calculate the thresholds to reduce the false negative rate and false positive rate. In order to calculate these thresholds, we collect 1000 images captured at the building entrance and input them to detector. Then we count the false negatives and false positives on these 1000 images and calculate the thresholds as follow,

$$\min_{\theta_1,\theta_2,\theta_3 \epsilon [0,1]} \frac{\sum_{i=1}^{N}(FN_i(\theta_1,\theta_2,\theta_3)+FP_i(\theta_1,\theta_2,\theta_3))}{N}, \quad (1)$$

where N is the number of test pictures; FN_i is the false negatives of the i picture and FP_i is the false positives of the i picture; $\theta_1, \theta_2, \theta_3$ are the threshold of P-Net, R-Net, O-Net, which is 0.5, 0.6, 0.6 in our system.

3.2 Face Tracking

It is challenging to detect faces in non-cooperative environments. Firstly, some people may be occluded by others who stand in front of them. Some people may wear facial ornaments such as eyeglass or mask. Secondly, it may be difficult to detect deformed faces such as lowing head. Finally, motion blur makes faces hard to detect in the video. Optimized MTCNN is not robust to detect multimodal faces. For example, a person face is detected when he walks straight to the camera. However, when he lowers his head or another person blocks part of his face, the face may not be detected. If the same person's face is detected at different times, the face at different times might be recognized as different identity. So a person may get two different labels, which is not allowed in attendance marking system. In this paper, we adapt tracking algorithm to help detect occluded, deformed and blurred faces. Therefore, a person face can be always detected in our environment. The detected images of a person constitute the tracklets. Our tracking algorithm improves not only the performance of detection, but also the performance of recognition. We recognize a person using detected face tracklets, which performs better on deformed faces. The result of tracking is shown as Figure 2.

Our tracker is designed by DSST[3] to track the object in a video efficiently. When a face is detected, the tracker keeps track of it. For every k_0 frames, the detector will be updated to correct the tracker. The detector assisted by the tracker ensures that the detected face will be tracked continuously even though it is occluded and deformed. Therefore, a person will be always detected and recognized as only one identity.

We use the proposed detector and tracker to process the video stream and get a series of face tracklets P^t, as defined in Eq. (2). When the IOU value between the tracking box and the detection bounding box is greater than 0.2, the detection bounding box replaces the tracking box to correct the tracker. If the IOU value between the detection bounding box and any tracking box is lower than 0.2, the detection bounding box is considered as another newcomer.

$$\boldsymbol{P^t = k(D^t \cup C^t) \cup (1-k) C^t}, \quad (2)$$

t indicates the frame number. P^t is the set of tracklets of frame number t; D^t, as defined in Eq. (3), is the set of detection bounding box of frame number t and C^t is the set of tracking bounding box of frame number t, which is defined in Eq.(4). Besides, k, defined in Eq. (5), represents whether to update the tracker.

$$\boldsymbol{D^t = \{d_j^t, j = 1, 2, \dots, M\}}, \quad (3)$$

M is the number of detection bounding box of frame number t; d_j^t is the detection bounding box number j of frame number t.

$$C^t = \{\alpha_i^t c_i^t, i = 1,2,\dots,N\}, \quad (4)$$

N is the number of tracking bounding box of frame number t; c_i^t is the detection bounding box number i of frame t; α_i^t indicates that whether to remove the tracking bounding box of number i in frame t, which is defined in Eq. (6).

$$k = \begin{cases} 1 & t = 1, 1+k_0, 1+2k_0 \dots \\ 0 & others \end{cases}, \quad (5)$$

$$\alpha_i^t = \begin{cases} 1 & max\{\mathrm{IOU}_{ij}, j = 1,2,\dots,M\} < 0.2, c_i^t \cap l = \emptyset \\ 0 & others \end{cases}, \quad (6)$$

l is the boundary of the images; IOU (Intersection over Union) is defined in Eq. (7).

$$\mathrm{IOU}_{ij} = \frac{area(c_i^t \cup d_j^t)}{area(c_i^t \cap d_j^t)}, \quad (7)$$

IOU_{ij} is the intersection over Union of d_j^t and c_i^t.

3.3 Face Recognition

Face recognition uses the detected faces as the input. In our face attendance marking system, Inception-ResNet-V1[12] is used to extract features from the detected faces and 512 dimensional feature vectors is obtained. However, face recognition is hard to achieve robust performance on the deformed faces in non-cooperative environments. In order to achieve robust recognition of deformed faces, we extract features of tracklets instead of a single image to realize face recognition. In our system, we extract the features of the first 20 frames in tracklets created by detection and tracking. The cosine distance between these features and face features in the multimodal face feature reference gallery is calculated to find the closest face feature in tracklets with the reference face feature. And we regard this feature corresponding label as the label of the whole tracklets obtained from the same face.

3.4 Face Feature Reference Gallery

According to the strategy we used, how to establish the face feature reference gallery will directly affect the performance and speed of face recognition. Some systems[2][4][8] established reference galleries which contain single face image per person. Although they have faster speed in recognition, their recognition performance is limited by the heavily deformed images which are quite different from the images in galleries. On the contrary, some reference galleries[9] contain too many face images per person, causing unnecessary redundancy and slower speed.

Figure 3. Face image examples of five different angles from three different people.

To balance the performance and speed of our system, we establish the multimodal face feature reference gallery. It contains five face images from five different view per person as shown in Figure 3. Meanwhile, our reference gallery stores image features extracted by CNN instead of images, which avoid the time and storage consumption on repeatedly extracting features.

4. EXPERIMENT RESULTS

4.1 Experiment Setting

Our face attendance marking system is fully implemented in tensorflow[1]. The system runs on the PC equipped with Intel® Core™ i7-6850K CPU with frequency of 3.60GHz, NVIDIA GeForce GTX 1080Ti GPU, 128GB RAM, and the system runs on Ubuntu16.04 64-bit operating system. The Logitech C270 camera is used to film video.

We create our own dataset to evaluate the attendance marking system. The dataset samples are shown in Figure 4. The dataset contains 8 videos, which is filmed at the building entrance. There are one or two people in each video. Each video, which is about 10 seconds long, contains face occlusion, deformation, blurring, etc. So, detecting and recognizing faces in videos is challenging.

4.2 Recognition Results

The system is evaluated on our own dataset. To show the effectiveness of our system, we design another system called "traditional system" which only uses detection and recognition to achieve attendance marking. In this traditional system, MTCNN is applied to detect faces and Inception-ResNet-V1 is used to extract face features. The face features reference gallery contains features from 10 different people. The results of the proposed system and traditional system are shown in Figure 5. (a) and (b) indicate that the proposed method can detect faces that traditional system can't. (c) and (d) show that our system has more robust recognition of deformed faces. (e) and (f) show that our system not only works well on blurred faces, but also on occluded faces. Traditional system is difficult to recognize blurred faces correctly. Our system can detect and recognize faces with severe occlusion, which is challenging to traditional system.

In order to prove our reference gallery is practical to increase the accuracy rate of recognition, two experiments are conducted on 1000 frames in "traditional system". In the first experiment, we take only one standard positive face image for per person as reference gallery. In the second experiment, we extract features from images of five different angles for per person as our reference gallery. The result is shown in Figure 6. As we can see, it is easy to get a false recognition result in the first experiment when the face is deformed. In the second experiment, our accuracy rate of recognition is relatively high so we don't need to collect more images for each person, which may cause computational burden. And we record average time consumption of each frame in two experiments, it takes 0.030 seconds in the first experiment and 0.015 seconds in the second. Thus it can be seen that our method achieves not only higher recognition rate but also higher speed.

An important aspect of face attendance marking system is that when the person enters a room, he will not be identified as two or more different identities at different time. We experiment on the dataset and the results are shown in Figure 7. From the experimental results, we can see that the method of traditional system identifies one person as more than one person. Beneficial from our tracklets and multimodal face feature reference gallery, we can achieve robust identification of a person. So, when a person enters a room, we only identify him as one individual, not multiple, which guarantees the accuracy of the attendance marking system.

5. CONCLUSION

This paper presents a real-time face attendance marking system used in non-cooperative environments. The method of the system has good robustness for detecting and recognizing deformed, occluded and blurred faces. The experiment results prove that our proposed method performs better on attendance marking in non-cooperative environments.

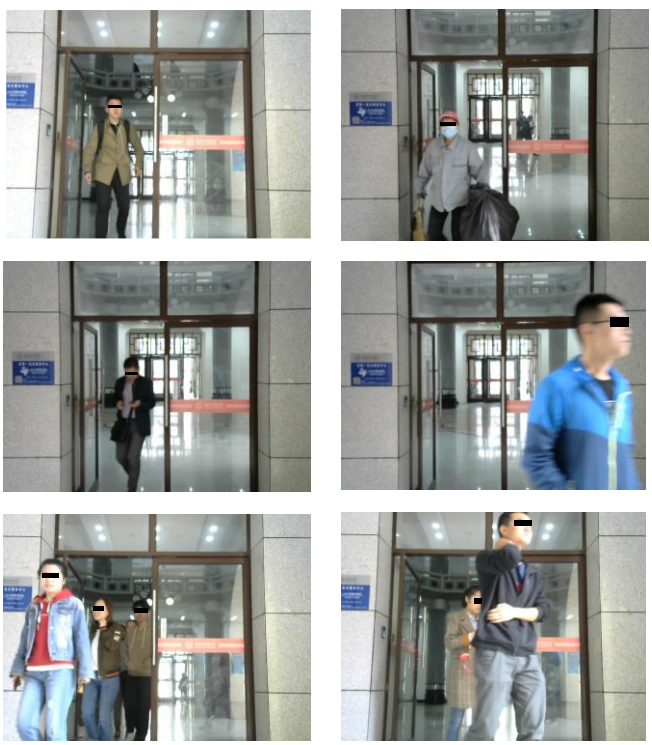


Figure 4. The dataset examples containing normal, blurred, occluded and deformed faces.

6. ACKNOWLEDGMENTS

This work is supported by the Key R&D Program—The Key Project of Shaanxi (Grant No. 707275177061)

7. REFERENCES

[1] Abadi M, Barham P, Chen J, et al. 2016. Tensorflow: a system for large-scale machine learning. *OSDI*.

[2] Chintalapati S, Raghunadh M V. 2013. Automated attendance management system based on face recognition algorithms. *ICCIC*.

[3] Danelljan, Martin, et al. 2014. Accurate scale estimation for robust visual tracking. *BMVC, Nottingham*, September 1-5.

[4] Kar N, Debbarma M K, Saha A, et al. 2012. Study of implementing automated attendance system using face recognition technique. *International Journal of computer and communication engineering*.

[5] Parkhi O M, Vedaldi A, Zisserman A. 2015. Deep face recognition. *BMVC*.

[6] Sajid M, Hussain R, Usman M. 2014. A conceptual model for automated attendance marking system using facial recognition. *ICDIM*.

[7] Schroff F, Kalenichenko D, Philbin J. 2015. Facenet: A unified embedding for face recognition and clustering. *CVPR*.

[8] Selvi, K. S., Chitrakala, P., & Jenitha, A. A. 2014. Face recognition based Attendance marking system. *International Journal of Computer Science and Mobile Computing*, *3*(2).

[9] Sharma, S., Gupta, T., & Kumar, R. 2018. Face Recognition in Real Time for Attendance Marking System. *International Journal of Scientific Research in Science and Technology*.

[10] Shriwastav, S., & Jain, D. C. 2016. A Review on Face Recognition Attendance System. *International Journal of Computer Applications*, *143*(8).

[11] Sun Y, Liang D, Wang X, et al. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873, 2015*.

[12] Szegedy C, Ioffe S, Vanhoucke V, et al. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI*.

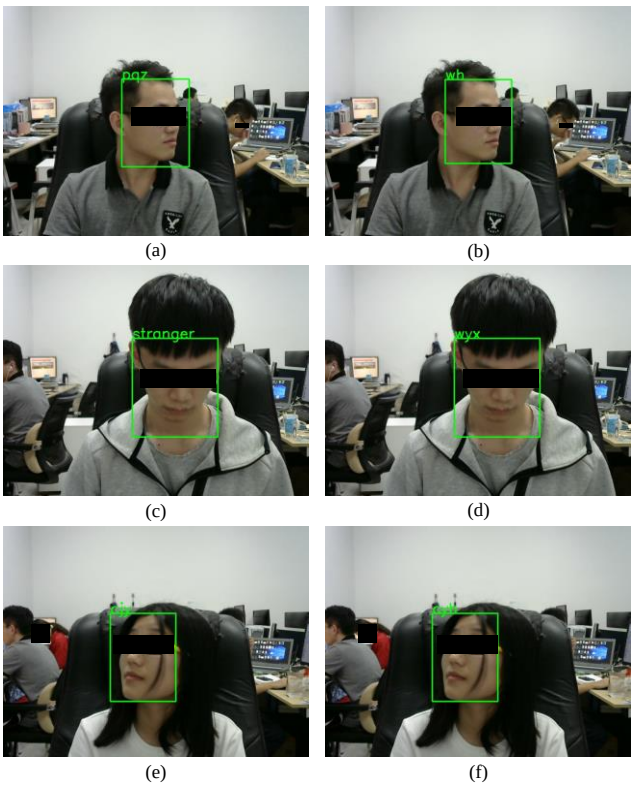
[13] Wen, Y., Zhang, K., Li, Z., & Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. *ECCV. Springer, Cham*.

[14] Yin, X., & Liu, X. 2018. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, *27(2), 964-975*.

[15] Zhang K, Zhang Z, Li Z, et al. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*.

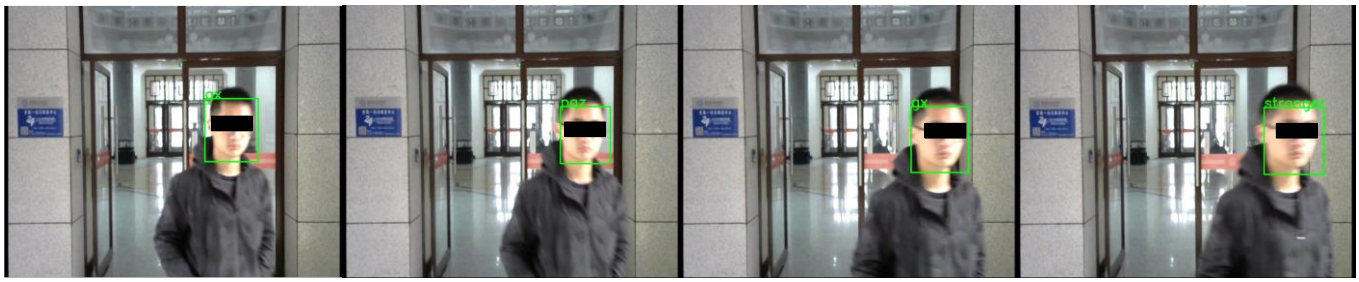
(a) (b) (c) (d) (e) (f)

Figure 5. (a) (c) (e) show the result of traditional system. (b) (d) (f) show the result of proposed system. The label in (a) (b) (c)(d)is "gx", the label in (e)(f) is "jk" and "wfy".



(a) (b) (c) (d) (e) (f)

Figure 6. The label of (a)(b) is "wh", (c)(d) is "wyx", (e)(f) is "cyh". As we can see, it is easier to get a false recognition in the first experiment(a)(c)(e) than it in the second experiment (b)(d)(f).



(a)

(b)

Figure 7. (a) is the recognition results of traditional system. (b) is the recognition results of proposed system. The identity produced by face recognition changes through the move of people in (a). The recognition results are robust to the people moving in (b).

Chapter 2: Target Detection

Top View Person Detection and Counting for Low Compute Embedded Platforms

Prashant Maheshwari
Capillary Technologies
prashant.m
@capillarytech.com

Doney Alex
Capillary Technologies
doney.alex
@capillarytech.com

Sumandeep Banerjee
Capillary Technologies
sumandeep.banerjee
@capillarytech.com

Saurav Behera
Capillary Technologies
saurav.behera
@capillarytech.com

Subrat Panda
Capillary Technologies
subrat.panda
@capillarytech.com

ABSTRACT

In this paper, we present an optimised approach for the top-view person detection and counting for low compute embedded platform. Earlier methods have used background subtraction for top-view detection which produces inaccurate results due to merging of blobs when there are multiple people in the frame. Several Deep Learning methods have been proposed for detection and tracking but they are computationally expensive to run on low compute device. We present an adaboost classifier for top-view detection and Kalman filter-Hungarian assignment for tracking which is optimised to give high accuracy on a low compute embedded platform. We achieved a counting accuracy of 97% in real-time (upto 40 FPS) on a Raspberry Pi3B. We also present a heuristic approach to handle false positives in real time through dynamic learning and unlearning of detections along with other optimisations in tracking.

CCS Concepts

•Computing methodologies→Computer vision problems Tracking • Computer systems organization→Real-time systems.

Keywords

Tracking, Real-time systems, Object detection, Multithreading

1. INTRODUCTION

People counting in various scenarios has drawn a lot of attention due to ever increasing use cases such as - surveillance, estimating the number of people in a large capacity building, business performance of any retail store through number of people visiting store. They rely on the extracted information from the detections and tracking - like the path traversed or the exact number of people crossing an area. Most of the research work has approached person detection using fronto-parallel or front view of person. However, it makes people counting a challenging task largely due to occlusion and depth perception which makes both detection and tracking inaccurate and difficult. There are numerous deep learning based approaches that have surfaced in recent times achieving state-of-the-art accuracy in both detection and tracking. These approaches are computationally expensive and requires specialised hardware to obtain results in real time. The top-view gives the advantage of almost no occlusion but the features available for detection are less compared to pedestrian or front-view and it also has perspective distortion. Therefore, detection in top-view is comparatively more challenging. The methods of segmenting foreground from background has been used for person detection in top-view [3] but it is susceptible to illumination changes, shadows and reflections. The reflections and shadows can be handled with the work by [19] and [39]. But, accurately splitting the merged blobs when there are multiple people in the frame in exact people count is a daunting task. Using stereo vision helps in splitting the blob more accurately [22], [28] but that requires extra hardware.

The people counting system has more impact if it is real-time and cost effective when it comes to mass installation. Therefore, in this paper we present a fast and resource efficient people counting algorithm for top-view monitored area which can run on low compute inexpensive embedded systems. The overhead mounted camera monitors a small area of few sq.ft, similar to that of office or shop entrances like in fig:1. To achieve reasonable accuracy and high detection FPS we overcome the following challenges : 1) to avoid occlusion we use top-view camera installation instead of pedestrian view. This slightly offloads the burden of accurate detection from our top-view detector. 2) We use a feature based top-view person detector based on Adaboost classifier which is robust to illumination and lighting changes and perspective distortion unlike background subtraction. BG-sub requires splitting blobs of people in close proximity, either heuristically which causes error in person count, or by estimating person size using stereo vision which requires extra hardware. Our detector does not have the above constraints. 3) We use Kalman Filter along with Hungarian assignment for tracking and we are also proposing numerous heuristic optimisations for better tracking, including dynamic handling of false positives through learning-unlearning of detections. 4) Lastly, we have used a multithreaded approach for our detection and tracking to achieve 40 FPS on Raspberry Pi. The paper is organized as follows: Section 2 review different solutions proposed in literature. Section 3 Proposed solution and Section 4 Experiments and Results.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301548

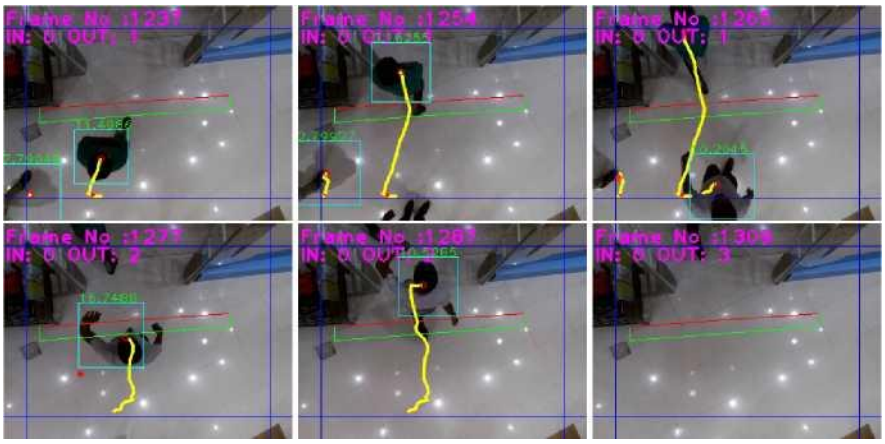


Figure 1. Fig a-f shows the detection, tracking and counting

2. RELATED WORK

Many people counting algorithms have been proposed and these approaches primarily vary with respect to the following factors: (i) Methods —detection or estimation, (ii) Person view —fronto-parallel or top view, (iii) Technology —Deep learning or machine learning. There are numerous people counting algorithms and approaches varying in terms of method, person view - front view or top view, and technology (deep learning or pure machine learning). A recent work that addresses our problem is people counting in dense crowd scenes [42], [20], [31]. In earlier approaches of pedestrian counting, person detection played an important role. They performed well as the numerous neural network architectures like [33], [34], [26], gives best of the accuracy for pedestrian detection even when occlusion occurs. But building a real time people counting system using neural networks is not feasible due to their slow performance on a low compute embedded platform such as Raspberry Pi 3B. Table 1 lists the time benchmarking of different architectures. These networks achieve less than 1 FPS which makes them unsuitable for tracking and hence for counting. However, there are other compute efficient algorithms for pedestrian detection which give good accuracy and speed. [9] and number of its modifications [44] perform well on pedestrian detection. Another notable work is by P Dollar [11], [10]. There also exist numerous people counting method based on pedestrian detection [40], [6], [25] and [4]. These methods are not very accurate as they fail to handle occlusion.

Previous work in top-view person detection [3], [36], and [24] have use background subtraction which did not produce satisfactory results as it was highly susceptible to illumination condition, change in lighting, and shadows. The variation in illumination changes creates a difference between the background pixel values and the learned background pixel values, therefore background gets detected as foreground blobs. These techniques are also very susceptible to outdoor scenarios where lighting conditions change based on weather. Similarly, the shadows also get detected as foreground blobs. Later works countered the effect of illumination, shadow and lighting using techniques like [19] and presented in [39], and [16]. Even after all these optimizations the blob detection using BG- subtract fails whenever two or more people are in close proximity in the frame producing a single fused blob and getting counted as one. Heuristic methods [32] based on human size can be used to split the blobs but such methods fail when people are carrying trolley or luggage/baggage with them. There are methods that use stereo vision to separate the fused blob of multiple people into exact count by estimating human size using depth information. There are other approaches where multiple cameras [43], [37] are used for people counting. But, the extra processing power required to process second camera feed and cost of multiple cameras makes the use of these methods undesirable. Apart from these, there are some bayesian based probabilistic approaches as suggested in [8] but these are designed to work with horizontal view, and hence fail when applied in overhead view. Similarly, Modification of HOG/SVM [2] - since only limited information is available in overhead view, relying only on HOG gives less accurate results.

Apart from detection, tracking those detection is the key in counting. Several methods exists for overhead tracking [13], [17] and [1] but these methods lack predictive nature in their approach which is key for robust system of counting. There are other single and multi object trackers like [30], [38], [7], [15] that gives state-of-the-art accuracy. However, the frame speed achieved from these methods is very minimal, MDNet and VITAL give 1 FPS and 1.5 FPS respectively with Tesla GPU. While, RDT, BACF gives 30-40 FPS but requires GPU. Achieving a real time tracking on low compute device would not be possible with such compute intensive algorithms. The tracking job needed to be computationally light and accurate as most of the computation power was already consumed by the detector. Kalman filter [23] along with Hungarian [29] assignment seems the most common method. Several complex methods for accurate tracking [41], [18] seems to be computationally expensive, hence not suitable for our use case. We used a simple Kalman tracking and Hungarian assignment combination with position and velocity as the parameters.

3. METHODOLOGY

The idea is to develop a simple system that would monitor a small area from overhead view and count the number of people passing through monitored area. With the recent success of neural networks in pedestrian detection and transfer learning, one possible approach is to re-train a fronto-parallel person detection networks for top-view detection. But deploying these computationally heavy networks on a portable low compute edge for a real-time application in cost effective manner is not a feasible task with today's hardware. One possible solution is to have a GPU based server which can run these algorithms in real time. But, that would require sending data through internet. For a

system monitoring an area and tracking each detection, it would require sending every captured frame to the server and therefore, require a lot of bandwidth and reliable up-time of network. This is not a reliable solution in cases such as latency and down time of network which can not be controlled for a real time application. The other method could be using Neural compute stick by Intel along with Raspberry Pi. But from that we could achieve 4.39 FPS [35] using SSD mobile net [21] which is not sufficient for real-time tracking. So is the case with neural network based tracker. Hence, we were required to develop a reliable detector and tracker algorithm which could be made to run on a small hardware device and can work as an offline standalone system. Detecting people in the frame using background subtraction by determining the foreground blob to be of single person or multiple people is a cumbersome task even to a human eye. In HOG-SVM approach the Histogram-Of-Gradients alone is not sufficient and more information is required for it to work on overhead view. We chose the Aggregated-Channel-Features based detector proposed by P.Dollar et al [12] because of the additional three color channels which helps capture head and partially visible face features while the remaining six channels for gradients in six directions models the physical structure of head and shoulders. After detection however a robust tracker is needed which should negate the effect of missed and false detections as explained in Figure 2. Kalman filter's ability to predict the next position of tracks without actual detection made tracking robust towards missed detections, while Hungarian assignment terminates the track initiated due to false positives. Further optimizations were made in detector and tracker to improve performance. We also discuss the multi-threaded implementation and how it boosts the frames per second throughtput of our algorithm.

The following subsections of paper shed light on optimization, multi-threaded implementation and people counting zone.

3.1 Overhead Detector

The overhead detector is based on the work done by P.Dollar et al on pedestrian detection using Aggregated Channel Features (ACF) with Adaboost Classifier. The overhead detector is trained on the images extracted from the videos taken from an overhead installed camera. To run it in real-time on low compute devices, optimizations were made on restricting detection scales and capturing video at optimal resolution. The resolution of input video is computed using average person size such that the person gets detected in first few scales. The average person size is a function of installation height and taken as input during device setup. For a camera installation height 'h' where h e H = [hi, h2, h3,..., hn}, there exist a scale 's' where s e S = [si, s2,..., si2} in which detection occurs (according to P.Dollar work these are 12). Corresponding to each scale there exist an image resolution r e R = [ri, r2,..., ri2}. For a person getting detected in scale say 's5', the computation done on scales from si - s4 is unnecessary. Therefore, the input video is recorded at a resolution corresponding to s5 such that s5 is the first scale of detection.

For a camera installation height 'h', the variation in person size in pixels is limited and is a function of person's height. Experimentally, we found that this can be covered in 4 scales of detector, therefore, saving computation of remaining 8 scales. We also found that as the resolution of image decreases - the false positive increases. Therefore, the two optimization improved the performance in terms of time and also reduced the occurrence of false positives.

3.2 Tracking

The actual In/Out count must be updated only when person enters/goes out of the view completely, we need an algorithm which is able to handle cases such as (i) people loitering in the frame should not be counted even if the person crosses the mid-line multiple times. (ii) false counts resulting from people changing their minds halfway even if they have crossed the mid-line should not be counted as they have not completed their sequence. This is especially true in the use case of retail people counter. (iii) No detector is 100% accurate. There are missed detections and false positives. Therefore, we propose a novel tracking logic which is be able to cater to all of the above situations and be robust enough to predict accurate count.

Missed detection causes premature termination of tracks, while false positives initiates false tracks and also affect the true tracks by latching (explained in figure 4. Fine tuning of kalman filter predictive feature based on displacement and velocity resolves the problem of missed detection and keeps the track from pre-mature termination. However, the problem of latching among tracks still persists due to:

1)improper detection at the edges of frame (partial view of person).

2)false positive. Latching can be understood from the following cases: 1) When the person goes out of frame, the track should end with no detection assignment, but instead, the track may be assigned to the false positive and the count is missed as shown in figure 4(i). 2) Similar situation occurs when missed detection happens, the track of true positive may be assigned to false positive and when person is detected again, a new track is initiated and the old track is lost to false positive as shown in figure 4(ii). This leads to inaccurate tracking and count miss. 3) Latching of false positive track to true positive creates false initiation point of track or worse, lead to count miss as shown in figure 4(iii). These detection and latching issues combined together creates lot of inaccuracy in tracking and the count. We present a novel approach, Figure 3, to handle false positive - dynamic learning and unlearning of false positives. We maintain the average confidence, number of missed frames and number of detected frames for every detection in tracker. If any detection crosses the threshold of number of detected frames (learning rate) then that detection is not passed to Kalman and Hungarian. This history is maintained for nearby area of detection instead of a single point to negate the effects of jitter. This solution can cause the problem of not passing the true positive to tracker for the learned region but that region is very small and can be handled by tracker. There were cases where a person standing in the frame is learned, this way we could end up with entire FOV turning into missed detection region. To differentiate between a learned true positive and false positive, we again refer to the properties of detection. A true positive, even though learned, will not be present for the entire duration, therefore, the unlearning of detection. If number of missed detection frames is greater than the threshold of learning rate, then that detection is unlearned.

At the edge of the frame, partial visibility of person gives jittery detection. The person moving out of frame has outward velocity and acceleration, but detected center remains constant due to gradually reducing detected partial rectangle size. The stationary center and jitter gives apparent deceleration in contrast to actual direction of movement. This leads to Kalman filter prediction in opposite direction. This estimation may cause the track to latch on the nearby detection, illustrated in figure 5(i). The information for

the two tracks is lost due to latching, first of the actual track and second of the new detection on which track latched on. To overcome this problem, we introduced the region based tracking. These regions at the corner provides the buffer to terminate tracks gracefully and act as virtual frame boundaries beyond which if detection happens then the corresponding track is terminated thereby preventing latching as illustrated in figure 5(ii). Therefore, we avoid the error prone edges.

3.3 Counting method

We divided our FOV in different regions as Red, Yellow, Green, Pink and Cyan showed in figure 6. The counting is done based on region of first detection and last true detection. We maintained the starting region and ending region for every track and when the track is terminated count is updated. Count increment cases are as follows: Red to Green - increase IN count, Red to Cyan - increase IN count, Green to Red - increase OUT count, Green to Pink - increase OUT count. While the yellow region is the buffer zone where no change of count takes place whether detection begins or ends. Pink region is virtual boundary towards exterior. Cyan is virtual boundary towards interior.

3.4 Real Time Implementation

The accuracy of the algorithm is heavily dependent on the frame rate of the system as tracking accuracy suffers as frame rate decreases. The fps achieved with a single threaded implementation was very low. The time taken to process a frame was also dependent on the number of people in the frame as all the weak classifiers of Adaboost has to run on each block for a true positive. Hence the fps falls even further when there are multiple people in the frame bringing it below the average fps. A single threaded implementation, after all the optimization could not achieve the required accuracy. Hence a multi-threaded implementation was required.

The algorithm was implemented with an asynchronous multithreaded design illustrated in figure 7. The input frames were queued into an input buffer queue. This made sure the system was working in a constant frame rate irrespective of whether there are people in the frame or not.

The detector part of the algorithm was heaviest (taking almost 80 to 90 of percent of the processing). The detection for a particular frame was independent of any previous or future frames. Hence multiple instances of detectors could work independently.

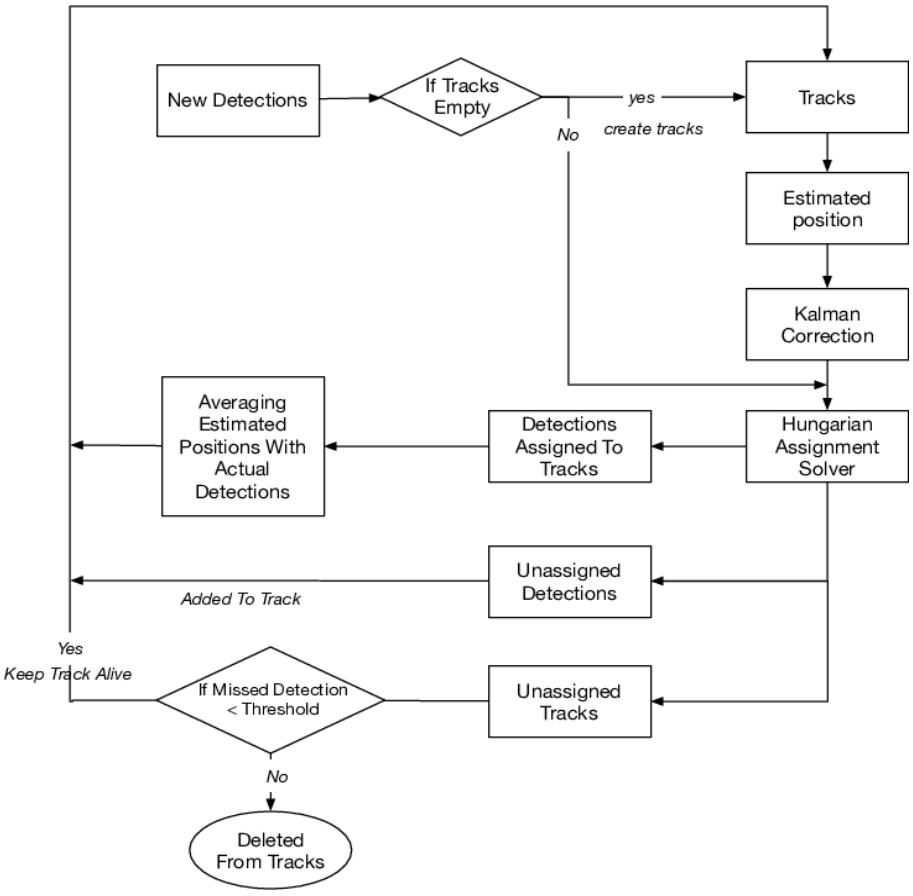


Figure 2. This figure illustrate the flow chart for Tracking algorithm

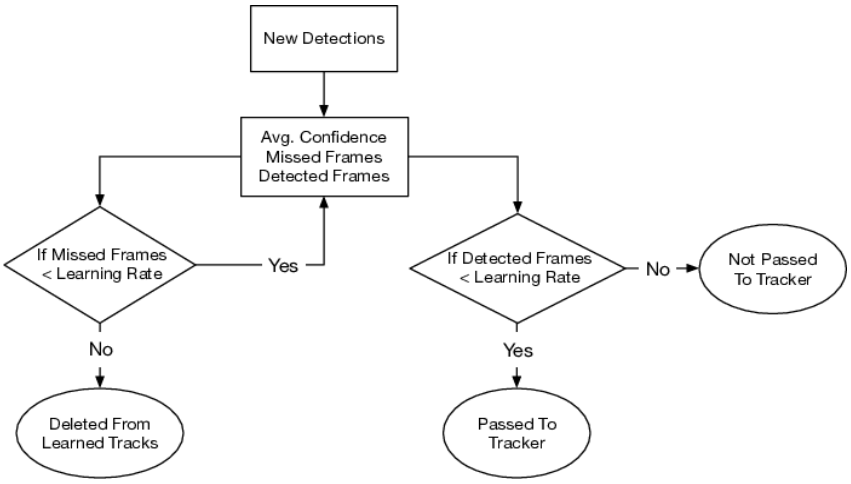


Figure 3. This figure illustrate the flow chart for following: (a) Learning/Unlearning of false positive

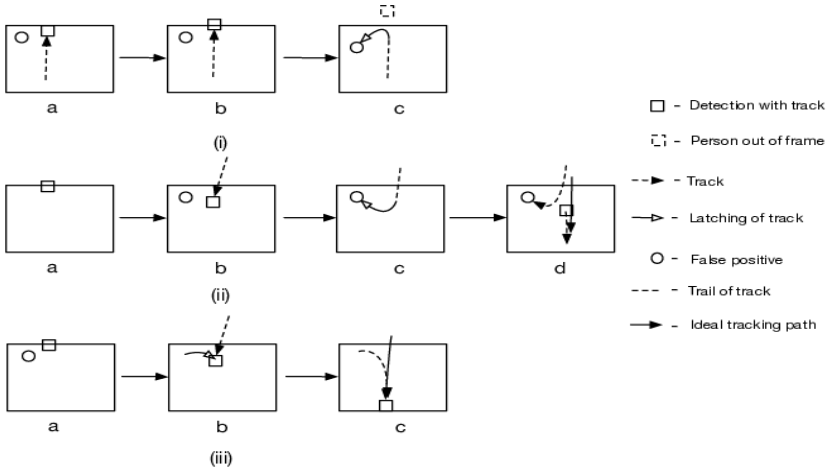


Figure 4. This illustrates Latching scenarios:® tracking jitter & deceleration at edge of the frame along with false positive causing latching. (ii) Appearance of false positive and missed detection, causing original track latching on to false positive & new track initiated for actual detection. (iii) Track initiated by false positive being latched on to true positive.

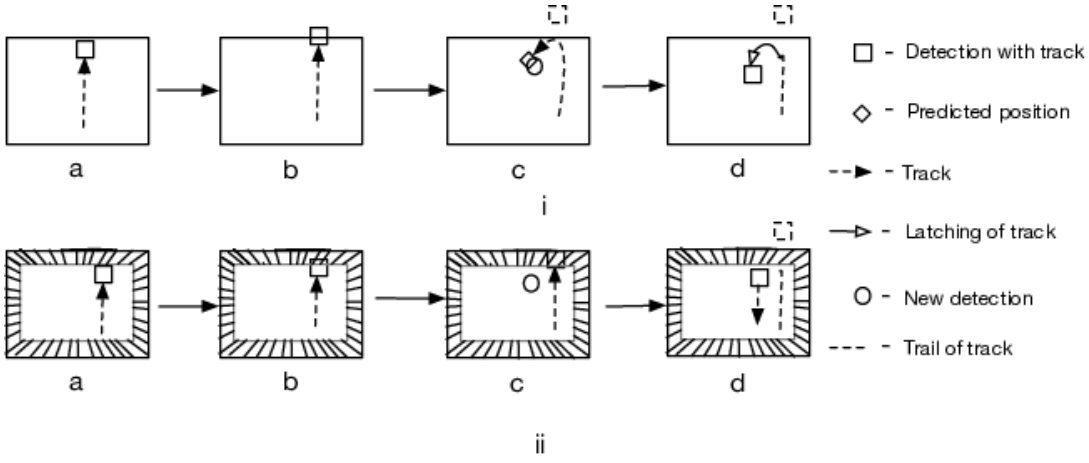


Figure 5. This illustrates the effect of buffer zone, i) In the figure, 'a' & 'b' represents tracks life, 'C' shows predicted position due to deceleration and also new detection, d shows that since predicted & actual detection is close by therefore latching. ii) The shaded region is the buffer zone providing cushion to terminate tracks gracefully with no false prediction in 'c' and in'd' a new track is initiated for new detection.

Figure 6: zonemap

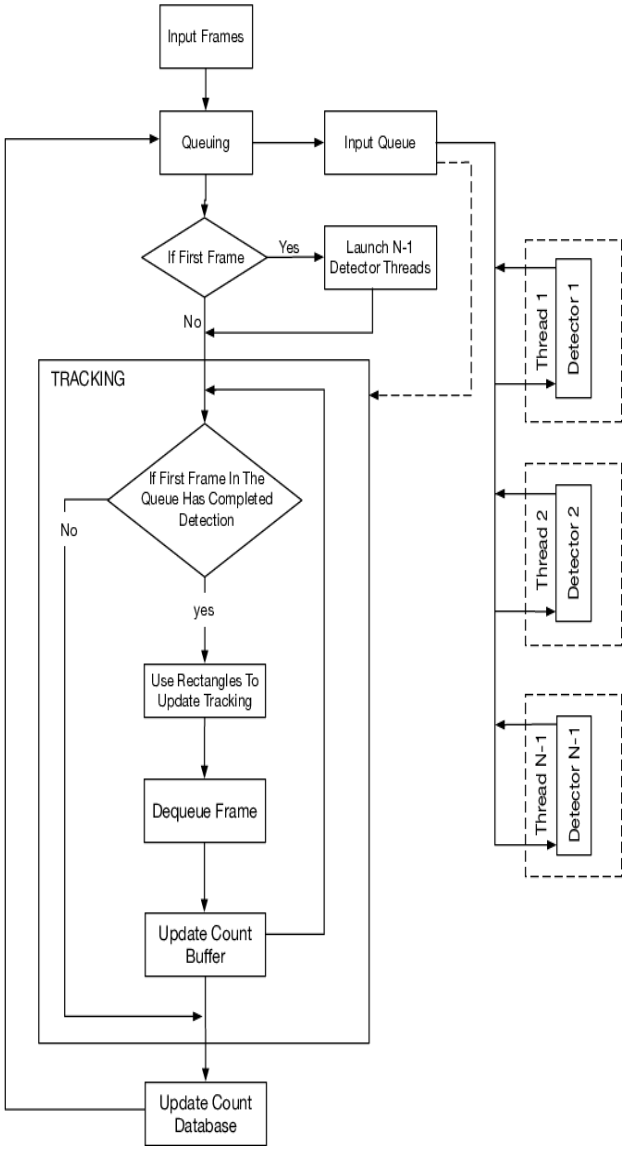


Figure 7. This figure illustrate the flow chart for multi-threaded design

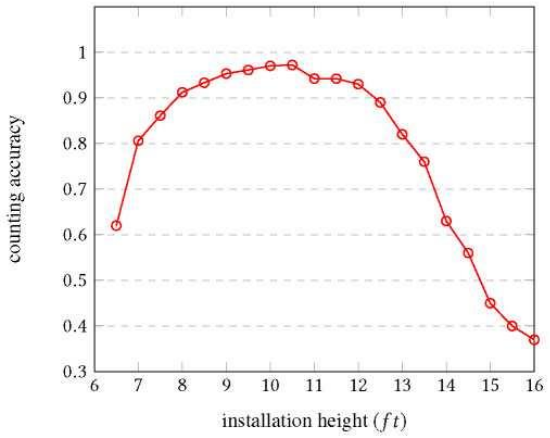


Figure 8. Counting accuracy for different installation height

The implementation was configured for N threads where N is an input parameter. Main thread does queuing of frames into the buffer and tracking job. The master thread launches N-1 threads, each performing detection on a particular frame. Once launched these threads run continuously until the exit is called. Each detector look for frames in the input queue in sequential order. It takes the oldest frame on which detection is not performed, does detection and puts it back into the queue along with detected rectangles, so that the order is maintained. Since the detection on a single frame is independent of other frames, the detectors can run asynchronously. The tracking needed to be sequential, hence there is only one instance of tracking which was performed by the main thread. The main thread after queuing an input frame, check whether the oldest frame in the queue has completed detection. If yes, it will use those rectangles to update tracking and dequeue the frame and looks for the next oldest frame. This continue until a frame on which detection is not completed is in the front of the queue.

The detector uses lot of memory for channels computation of Adaboost Classifier. Allocation and de-allocation of the memory repeatedly was adversely affecting the performance. Therefore, a custom memory management technique was implemented in which a chunk of memory was allocated once in the beginning and was reused for each frame. Similar technique used for input frame queue.

4. EXPERIMENT AND RESULTS

The state-of-the-art methods are slow on low compute-power devices. Doing real time counting is not feasible on such devices.

The novelty of the paper is the proposed methodology which gave an accuracy of 97% on device like R-Pi 3B in real time (i.e.20FPS). In this section we demonstrate the step by step improvement in accuracy due to algorithmic optimisations. The comparative study of our method could not be done with state of the art methods in terms of accuracy due to two reasons. 1) State of the art accuracy are achieved by Deep Learning based solutions. These accuracy could not be achieved in real time (20 FPS) on devices like R-Pi due to low frame rate. The FPS achieved by DL methods is mentioned in table 1. Such poor FPS discouraged us to proceed further with it. 2) The described algorithm is optimised for the presented overhead view as shown in 1. There is no dataset publicly available which has the same view. Several datasets [5], [27], [14], [12] are available for overhead view but they are not perfectly overhead view. They have an angled view where partial of face/side of head can be seen. The performance on two datasets could not be compared due to their different nature and therefore, our custom datasets (e.g in Fig:1) was used for accuracy computation and benchmarking. The accuracy is computed in two folds - algorithmic accuracy and real time accuracy on device. We created our own data-set by recording videos from overhead in different ambient conditions covering the following use cases. 1) Different installation height, 2) Changing natural lighting condition, 3) Artificial lighting, 4) Different flooring.

The implementation of detecting person in static background using background subtraction along with shadow detection produced good results under experimental setup giving an accuracy of 78%. However, when the algorithm is subjected to illumination and lighting changes, counting accuracy fell below 50%.

Our detector is robust and efficient towards sudden lighting, illumination, shadow changes such as switching On/Off light in the middle and people coming in groups. The optimization made in detector and tracker along with parallel threaded implementation gave a bi-directional counting accuracy of 97% at 20FPS. A typical view of the detection and tracking is shown in figure 1. The counting accuracy is computed videos of 25580 minutes with total bi-directional count of more than 11946. Ground truth is maintained by manually counting the number of people in videos. Criteria for a valid count is —

1) Person should be completely visible in the frame,

2) Person should leave the frame completely

3) If person enters and leaves the frame from same side then count is not incremented

4) If person walks side-ways after entering the frame then that person is not counted.

For single threaded approach, tracking suffered due to frequent missing of frames caused by compute bottleneck and hence dip in accuracy. Therefore, we took multi-threaded approach and present the performance results in 3. We also installed our devices on 110 public locations with different ambient settings and at different installation height to assess the real time performance. We did manual counting on these locations and present the result in figure: 8. We compare results of Background subtraction with our implementation in terms of effect on counting accuracy with the suggested improvement in detector & tracking. The improvement in accuracy after each step is shown in Table 2.

5. CONCLUSION

In this paper, we presented a state-of-the-art system for real time people counting. We suggested a novel approach of using Overhead person detector along with multiple optimization to improve performance and compute time. We also made numerous improvements in tracker and suggested a novel approach of handling false positive by dynamic learning/unlearning. We achieved an accuracy of 97% in real time (20FPS) for a low computing embedded device. For a very low installation height of camera the accuracy decreases. Future work can be done on improving accuracy for low height installations.

Table 2: Accuracy variation with various improvements in the Detector & Tracker. In the table, PD - person detector, BG-Sub - Background subtraction, OHD - Our Overhead person detector, KFHA - Kalman Filter and Hungarian assignment, IRFSD - Input resolution is the first scale of detection, BZ - Blue zone, SRD - Scale restriction in detector, FPL/U - False positive learning/unlearning, Accuracy - Bi-directional Counting Accuracy (in %)

Incremental changes in PD	Incremental changes in Tracker	Accuracy
BG-Sub	KFHA	40.9
OHD	KFHA	61.2
OHD + IRFSD	KFHA	70.8
OHD + IRFSD + SRD	KFHA + BZ	84.7
OHD + IRFSD + SRD	KFHA + BZ + FPL/U	97.2

Table 3: Multi-threaded vs single threaded (computed at frame rate of 20 FPS)

Approach	Bi-directional Counting Accuracy (in %)
Single threaded	73
Multi-threaded	97.2

6. REFERENCES

[1] Jonathan Owens A, Andrew Hunter B, and Eric Fletcher A. [n. d.]. *A Fast Model-Free Morphology- Based Object Tracking Algorithm.* ([n. d.]).

[2] I. Ahmed and J. N. Carter. 2012. A robust person detector for overhead views. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. 1483-1486.

[3] J. Barandiaran, B. Murguia, and F. Boto. 2008. *Real-Time People Counting Using Multiple Lines.* In 2008 Ninth International Workshop on Image Analysis for Multimedia

Interactive Services. 159-162. https://doi.org/10.1109/WIAMIS.2008. 27

[4] Tudor Barbu. 2014. *Pedestrian detection and tracking using temporal differencing and HOG features.* Computers & Electrical Engineering 40, 4 (2014), 1072 - 1079. https://doi.org/10.1016/j.compeleceng.2013.12.004

[5] A. B. Chan and N. Vasconcelos. 2008. *Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures*. IEEE Transactions on Pattern Analysis andMachine Intelligence 30, 5 (May 2008), 909-926. https://doi.org/10.1109/ TPAMI.2007.70738

[6] C. C. Chen, H. H. Lin, and O. T. C. Chen. 2011. *Tracking and counting people in visual surveillance systems.* In 2011 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP). 1425-1428. https://doi.org/10.1109/ ICASSP.2011.5946681

[7] Janghoon Choi, Junseok Kwon, and Kyoung Mu Lee. 2017. *Visual Tracking by Reinforced DecisionMaking.* CoRR abs/1702.06291 (2017). arXiv:1702.06291 http://arxiv.org/abs/1702.06291

[8] I. Cohen, A. Garg, and T. S. Huang. 2000. *Vision-based overhead view person recognition.* In Proceedings15th International Conference on Pattern Recognition. ICPR-2000, Vol. 1. 1119-1124 vol.1. https://doi.org/10.1109/ICPR.2000.905668

[9] N. Dalal and B. Triggs. 2005. *Histograms of oriented gradients for human detection*. 1 (June2005),886-893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[10] Piotr Dollar, Serge Belongie, and Pietro Perona. 2010. *The FastestPedestrian Detector in the West.* In Proceedings ofthe British Machine Vision Conference. BMVA Press, 68.1-68.11. doi:10.5244/C.24.68.

[11]Piotr Dollar, Zhuowen Tu, Pietro Perona, and Serge Belongie. 2009. *Integral Channel Features.* In Proc. BMVC. 91.1-91.11. doi:10.5244/C.23.91.

[12] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. 2012. *Pedestrian Detection: An Evaluation of the State of the Art. PAMI 34 (2012).* S. Duffner and C. Garcia. 2013. PixelTrack: *A Fast Adaptive Algorithm for Tracking Non-rigid Objects.* In 2013 IEEE International Conference on Computer

[13] Vision. 2480-2487. https://doi.org/10.1109/ICCV.2013.308

[14] M. Enzweiler and D. M. Gavrila. 2009. *Monocular Pedestrian Detection: Survey and Experiments.* IEEE Transactions on Pattern Analysis andMachine Intelligence 31, 12 (Dec 2009), 2179-2195. https://doi.org/10.1109/TPAMI.2008.260

[15] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. 2017. *Learning Background-Aware Correlation Filters for Visual Tracking.* CoRR abs/1703.04590 (2017). arXiv:1703.04590 http://arxiv.org/abs/1703.04590

[16] J. Garcia, A. Gardel, I. Bravo, J. L. LAqzaro, M. Martinez, and D. Rodriguez. 2013. *Directional People Counter Based on Head Tracking.* IEEE Transactions on Industrial Electronics 60, 9 (Sept 2013), 3991-4000. https: //doi.org/10.1109/TIE.2012.2206330

[17] A.B. Godbehere, A. Matsukawa, and K. Goldberg. 2012. *Visual tracking ofhuman visitors under variable-lighting conditions for a responsive audio art installation.* In 2012 American Control Conference (ACC). 4305-4312. https://doi.org/10.1109/ ACC.2012.6315174

[18] Joao F. Henriques, Rui Caseiro, and Jorge Batista. 2011. *Globally Optimal Solution to Multi-object Tracking with Merged Measurements.* In Proceedings of the 2011 International Conference on Computer Vision (ICCV '11). IEEE Computer Society, Washington, DC, USA, 2470-2477. https://doi.org/10.1109/ ICCV.2011.6126532

[19] Thanarat Horprasert, David Harwood, and Larry S. Davis. 1999. *A statistical approach for real-time robust background subtraction and shadow detection.* 1-19.

[20] Yaocong Hu, Huan Chang, Fudong Nian, Yan Wang, and Teng Li. 2016. *Dense crowd counting from still images with convolutional neural networks.* Journal of Visual Communication and Image Representation 38 (2016), 530 - 539. https: //doi.org/10.1016/j.jvcir.2016.03.021

[21] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and KevinMurphy. 2016. *Speed/accuracytrade-offs for modern convolutional object detectors.* CoRR abs/1611.10012 (2016). arXiv:1611.10012http://arxiv.org/abs/ 1611.10012

[22] Xiaoyn Huang, Liyuan Li, and T. Sim. 2004. *Stereo-based human head detection from crowd scenes*. In 2004 International Conference on Image Processing, 2004. ICIP '04., Vol. 2. 1353-1356 Vol.2. https://doi.org/10.1109/ICIP.2004.1419750

[23] Rudolph Emil Kalman. 1960. *A New Approach to Linear Filtering and Prediction Problems.* Transactions ofthe ASME-Journal of Basic Engineering 82, Series D (1960), 35-45.

[24] Damien Lefloch. [n. d.]. *Real-Time People Counting system using Video Camera.*

[25] J. Li, L. Huang, and C. Liu. 2011. *Robust people counting in video surveillance: Dataset and system.* In 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 54-59. https://doi.org/10.1109/ AVSS.2011.6027294

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: *Single Shot MultiBox Detector*. In Computer Vision - ECCV 2016, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 21-37.

[27]Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. 2013. *Crowd Counting and Profiling: Methodology and Evaluation.* Springer New York, New York, NY, 347-382. https://doi.org/10.1007/978-1-4614-8483-7_14

[28] Ruijiang Luo and Yan Guo. 2001. *Real-time stereo tracking of multiple moving heads.* In Proceedings IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. 55-60. https: //doi.org/10.1109/RATFG.2001.938910

[29] James Munkres. 1957. *ALGORITHMS FOR THE ASSIGNMENT AND TRANSPORTATION PROBLEMS.* (1957).

[30] Hyeonseob Nam and Bohyung Han. 2015. *Learning Multi-Domain Convolutional Neural Networks for Visual Tracking.*

CoRR abs/1510.07945 (2015). arXiv:1510.07945 http://arxiv.org/abs/1510.07945

[31] D. Onoro Rubio and R. J. Ldpez-Sastre. 2016. *Towards perspective-free object counting with deep learning.* In ECCV.

[32] Sangho Park and J. K. Aggarwal. 2000. *Head segmentation and head orientation in 3D space for pose estimation of multiple people.* In 4th IEEE Southwest Symposium on Image Analysis and Interpretation. 192-196. https: //doi.org/10.1109/IAI.2000.839598

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. *You Only Look Once: Unified, Real-Time Object Detection.* In 2016 IEEE Conference on Computer Vision and PatternRecognition (CVPR). 779-788. https://doi.org/10.1109/CVPR. 2016.91

[34] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* CoRR abs/1506.01497 (2015). arXiv:1506.01497 http://arxiv.org/abs/1506.01497

[35] Adrian Rosebrock. 2018. *Real-time object detection on the Raspberry Pi with the Movidius NCS*. (2018). https://www.pyimagesearch.com/2018/02/19/ real- time- object- detection- on-the- raspberry-pi- with- the- movidius-ncs/

[36] A. K. Sahoo, S. Patnaik, P. K. Biswal, A. K. Sahani, and P. B. Mohanta. 2013. *An efficient algorithm for human tracking in visual surveillance system.* In 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013). 125-130. https://doi.org/10.1109/ICIIP.2013.6707568

[37] Thiago T. Santos and Carlos H. Morimoto. 2011. *Multiple camera people detection and tracking using support integration.* Pattern Recognition Letters 32,1 (2011), 47 -55. https://doi.org/10.1016Zj.patrec.2010.05.016 ImageProcessing, Computer Vision and Pattern Recognition in Latin America.

[38] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson W. H. Lau, and Ming-Hsuan Yang. 2018. *VITAL: VIsual Tracking via Adversarial Learning.* CoRR abs/1804.04273 (2018). arXiv:1804.04273 http://arxiv.org/abs/1804.04273

[39] Jae won Kim, Kang sun Choi, Byeong doo Choi, and Sungjea Ko. 2002. S.-J.: *Real-time Vision-based People Counting System for the Security Door.* In: Proc. of 2002 International Technical Conference On Circuits Systems Computers and Communications, Phuket.

[40] Huazhong Xu, Pei Lv, and Lei Meng. 2010. *A people counting system based on head-shoulder detection and tracking in surveillance video.* In 2010 International Conference On Computer Design and Applications, Vol. 1. V1-394-V1-398. https://doi.org/10.1109/ICCDA.2010.5540833

[41] A. L. Yussiff, S. P. Yong, and B. B. Baharudin. 2014. *Parallel Kalman filter-based multi-human tracking in surveillance video.* In 2014 International Conference on Computer and Information Sciences (ICCOINS). 1-6. https://doi.org/10.1109/ ICCOINS.2014.6868359

[42] Cong Zhang, Hongsheng Li, X. Wang, and Xiaokang Yang. 2015. *Cross-scene crowd counting via deep convolutional neural networks*. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 833-841. https://doi.org/ 10.1109/CVPR.2015.7298684

[43] ZhognchuanZhang andF. Cohen. 2013. *3D pedestrian tracking based on overhead cameras.* In 2013 Seventh International Conference on Distributed Smart Cameras (ICDSC). 1-6. https://doi.org/10.1109/ICDSC.2013.6778235

[44] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and S. Avidan. 2006. *Fast Human Detection Using a Cascade of Histograms of Oriented Gradients.* In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06),Vol. 2. 1491-1498. https://doi.org/10.1109/CVPR.2006.119

Image Edge Detection using Fractional Order Differential Calculus

S. Gupta[1], A. Bhardwaj[2], R. Sharma[3], P. Varshney[4], S. Srivastava[5]
Division of ICE, Netaji Subhas University of Technology
Sector 3, Dwarka, New Delhi 110078,India
sangeeta22011@gmail.com[1]; ananditab.ic@nsit.net.in[2]; rajneesh496@gmail.com[3];
pragya.varshney1@gmail.com[4]; smriti.nsit@gmail.com[5]

ABSTRACT
Edge detection is a field of signal processing where the signal is an image. Edges contain most of the information making edge detection a very important image segmentation technique. Traditional edge detection uses integral order differentiation, the results of which are not satisfactory. Often, the edges are missing, fragmented or with false edges. In our proposed work, edge detection has been performed using fractional order calculus to overcome these drawbacks. Edges and noise, both are high frequency components and the presence of noise in an image can lead to erroneous results. Therefore, image denoising is performed before edge detection operation for a noisy test image.

CCS Concepts
Hardware→Communication hardware, interfaces and storage→Signal processing systems→Noise reduction.

Keywords
Fractional order Signal Processing; Edge Detection; Differential Mask Operator.

1. INTRODUCTION
When edge detection operation is applied to an image, only the edges, i.e., the boundary of various objects in that image are retained and the rest of the components of the image are removed. Therefore, edge detection operation is also used when image (data) compression is required. Edges contain most of the information of the image [1]. When the colour or brightness information of the image is not required, edge detection operation is used as it drastically reduces the size or storage space occupied by the image. In some cases, the clarity of the image increases as other parameters of the image are removed. In the area of fault diagnostics, edge detection operation can be utilised as most of the defects or faults in a given component are more visible if only the edges of the image are present.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301507

To implement edge detection, a variety of software's are available. Generally, a differential mask is created [2-5]. The mask is convolved with the test image. The result of the convolution of the test image with mask produces edges of the image. In the traditional methods, these masks are created using integral order calculus. However, traditional edge detection may not produce ideal edges. These edges maybe fragmented, false edges might appear and desired edge segments might be missing.

In a previous method, WANG [1] utilised wavelet fractional differential theory for creating the differential mask to carry out edge detection [6]. Fractional-order Signal Processing enhances the signal's middle and high frequency components. It also retains the low frequency components of the signal [7]. Our proposed work uses a fractional order differential mask. The texture information of the test image is enhanced during edge detection in our work. In addition, the smooth texture information is also retained. The layout of this paper is as follows: Section 2 gives the literature review and research methodology. Section 3 covers the details of edge detection and mask operator and section 4 covers fractional order differential mask. In section 5, our proposed procedure for fractional order edge detection is presented and results discussed. Section 6 concludes the paper.

2. LITERATURE REVIEW AND RESEARCH METHODOLOGY
Traditional edge detection operators include Prewitt, Sobel, Kirsch, Canny etc. All the traditional edge detection operators utilise integral order calculus for creating the differential mask operator. The results obtained with these operators are not satisfactory [2-5]. Due to these shortcomings, fractional order calculus was utilised in the field of image edge detection [8-9].In this paper, edge detection is performed using fractional order differential mask operator. Our proposed work modifies the fractional order of the mask, thereby including all the edges along with the texture information of the test image. It is an improvement over the traditional edge detection methods as all the edges appear, the edges are continuous and no false edges appear.

In the proposed work, the image under consideration is converted into a grey scale image. A fractional order differential mask is applied to this grey scale image, which results in effective edge detection. A fractional order differential mask is constructed prior to edge detection. The process is carried out for different values of the fractional order and an ideal value is selected on the basis of the quality of the image produced during edge detection.

2.1 Edge Detection and Mask Operator

Edge detection is the process of identifying sudden sharp brightness changes or discontinuity, also called edge in an image [10]. It is important to detect edges because they contain most of the information of an image. Edges also indicate the shape of the objects present in the image. It is an important tool in the area of feature detection and feature extraction. The edges refer to a surrounding pixel grey level change which exists between the target and background, the target and the target, the region and the region, the base and the base [11]. When edge detection operator is applied to an image, connected curves are observed which indicate discontinuities in image brightness. In addition to it, discontinuities in image brightness may imply:

(i) Discontinuities in surface orientation.
(ii) Discontinuities in depth.
(iii) Discontinuities in scene illumination.
(iv) Discontinuities in material properties.

As a result, the output of the edge detection operator might indicate variation in surface orientation, depth, and scene illumination or material properties. Using edge detection operation to an image, the amount of data that needs to be processed reduces drastically, information in the image which is not very relevant is filtered out and the information which is important and provides structural properties of the image are retained. However, edge detection of moderately complex images captured from the real world, may not produce ideal edges. These edges maybe fragmented, i.e., edges might not be continuous, false edges might appear and desired edge segments might be missing.

Differential operation is widely used in the field of signal processing. There are various approaches to carry out edge detection operation. The edge detection method may be classified as:

2.1.1 Search based Methods

Search based methods utilize first order derivative like gradient magnitude to compute the strength of the edge and then utilise gradient direction in order to compute the local orientation of the edges. Gradient algorithm is used to detect edges for a first order derivative expression of an image.

2.1.2 Zero Crossing Methods

For zero crossing methods, second order derivative of an image is computed. Zero crossings are determined in the second order derivative expressions for finding edges. Laplace operator is used to detect edges for a second order derivative expression of an image (Figure 1).

Classical edge detection is carried out by the use of various differential operators like Roberts, Prewitt, Sobel, log, Canny etc.

2.1.3 Mask Operator

To carry out edge detection, a mask is convolved with a test image. A mask is a two-dimensional signal which is used for filtering the image. Convolution of the image with the mask is used for the following image processing operations- blurring of image, edge detection, noise reduction in an image and sharpening of an image. The order of the mask can be 1x1, 3x3, 5x5 or 7x7. The order is always odd because, for carrying out convolution, the middle of the mask is required. The convolution of the image and the mask is performed as follows:

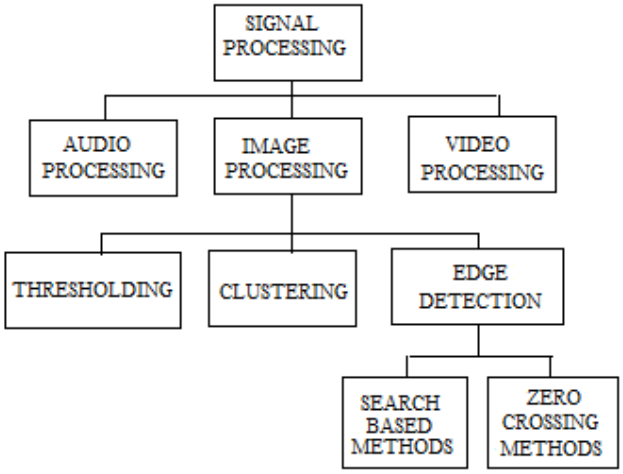


Figure 1. Flowchart depicting classification in Signal Processing

(i) The Mask is flipped vertically as well as horizontally.
(ii) The mask is moved on the image. The corresponding elements are multiplied and then added for a particular mask image position.
(iii) The process is repeated for all the pixel values.

An example for the convolution of an image with a mask occurs in following manner:

(i) Let the Mask and Image be

1	4	7
2	5	8
3	6	9

A	B	C
D	E	F
G	H	I

(ii) The Mask is flipped horizontally and then vertically

7	4	1
8	5	2
9	6	3

9	6	3
8	5	2
7	4	1

(iii) The middle coefficient of a mask is placed on a pixel of the image. Now the horizontally and vertically flipped mask is convolved with the image

9	6	3	
8	5A	2B	C
7	4D	1E	F
	G	H	I

$$Value\ of\ the\ First\ Pixel = 5*A+2*B+4*D+1*E \quad (1)$$

As the middle position of the mask is moved through every pixel of the image, the corresponding pixel values will be obtained. The edges are determined by the calculation of differences between the intensities of pixels of an image as they represent sudden brightness or intensity change. The masks which are commonly used for edge detection are Prewitt operator, Robinson compass operator, Sobel operator, etc. Since, images are nothing but a signal, and the changes in the signals are obtained using differentiation operation, therefore, these masks are also known as derivative masks.

The properties of derivative masks are as follows:

(i) The Mask coefficients must be present in opposite signs.
(ii) The sum of all the mask coefficients must be zero.
(iii) Edge detection is more efficient, if more weights are present in the mask.

3. FRACTIONAL ORDERDIFFERENTIAL MASK

In this work, edge detection model is obtained using fractional differential operator. G-L definition of the fractional order calculus is used as in [12]. The type of integer *k* order is generalized to arbitrary order operator D_v, where v is a real number.

$$\left(D_v \hat{f}\right)(\omega) = (i\omega)^v \hat{f}(\omega) = \hat{d}_v(\omega).\hat{f}(\omega) \qquad (2)$$

A Meta signal, $f(t)$ or differential expression of fractional order is derived:

$$\frac{d^v f(t)}{dt^v} \approx f(t) + (-v) f(t-1) + \frac{(-v)(-v+1)}{2} f(t-2) + \ldots + \frac{\Gamma(-v+1)}{n!\Gamma(-v+n+1)} f(t-n) \qquad (3)$$

$$\frac{\partial^v f(x,y)}{\partial x^v} \approx f(x,y) + (-v) f(x-1,y) + \frac{(-v)(-v+1)}{2} f(x-2,y) + \ldots + \frac{\Gamma(-v+1)}{n!\Gamma(-v+n+1)} f(x-n,y) \qquad (4)$$

$$\frac{\partial^v f(x,y)}{\partial y^v} \approx f(x,y) + (-v) f(x,y-1) + \frac{(-v)(-v+1)}{2} f(x,y-2) + \ldots + \frac{\Gamma(-v+1)}{n!\Gamma(-v+n+1)} f(x,y-n) \qquad (5)$$

The three coefficients $1,\ v,\ (-v)*(-v+1)/2$ are obtained.

The three coefficients obtained replace the corresponding position of zero order in all zero squares and a fractional differential mask operator is constructed on a two dimensional image. The fractional mask is obtained in the following directions:

(i) X direction
(ii) Y direction
(iii) Negative extraction
(iv) Negative Y direction
(v) Right lower diagonal direction
(vi) Right upper diagonal direction
(vii) Left lower diagonal direction
(viii) Left upper diagonal redirection

The corresponding coefficients obtained in all the eight directions are then added to obtain final mask template (Figure 2(a) to (i)) [1].

The Physical meaning of fractional differential signal in terms of signal processing is the generalization of amplitude phase modulation. As the frequency component or the differential order varies, amplitude of the signal varies and phase is the generalized Hilbert matrix of the frequency.

4. SIMULATION AND RESULTS

Fractional operator template (Figure 2) provides the final mask which acts as the edge detector. Fractional mask template is used for every pixel of image convolution. As a result, gradient of a point and the corresponding gradient amplitude is also obtained. Thus edge detection of the image is obtained. Fractional order Signal Processing results are better than traditional edge detection methods as all the desired edges were detected, the texture information of the image was enhanced and the smooth texture region of the image was preserved. The best order of fractional mask obtained in the proposed work is 0.7. If the differential order is low, the enhancement of texture of the image will reduce and there will be certain interference in the edge detection of the image. The results (Lena image) are shown in Figures 3(a) to (g).

5. CONCLUSIONS

Our proposed work successfully detects the edges of the test image by the application of fractional order signal processing. We address several shortcomings as the texture information of the image is enhanced and complete and continuous edges are obtained. The texture information of the smooth region is also preserved.The applications of the proposed work, in future, may include processing remote sensing images and medical imaging.

6. REFERENCES

[1] Wang, Z., Su, J. and Zhang, P. 2016. Image edge detection algorithm based on wavelet fractional differential theory. In *Proceedings of the35th Chinese Control Conference (CCC), 2016* (10407-10411). IEEE.

[2] Prewitt, J.M.S.1970, *Object Enhancement and Extraction Picture Processing and Psychopictorics*. B. Lipkin and A. Rosenfeld, eds., New York: Academic, 75-149.

[3] Hou, J., Ye, J.H. and Li, S.S. 2007. Application of Canny Combining and Wavelet Transform in the Bound of Step-Structure Edge Detection. In *Proceedings of the IEEE International Conference on Wavelet Analysis and Pattern Recognition*, 4, ICWAPR'07, 1635-1637.

[4] Gao, W., Zhang, X., Yang, L. and Liu, H. 2010. An improved Sobel edge detection. In *Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology*, *5*, (July 2010), ICCSIT, 67-71.

[5] Canny, J.1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 679-698.

[6] Pu, Y.F., Yuan, X., Liao, K., Chen, Z.L. and Zhou, J.L. 2005. Five Numerical Algorithms of Fractional Calculus Applied in Modern Signal Analyzing and Processing. *Journal Sichuan University Engineering Science Edition*,37, 5, 118-124.

[7] Pu, Y.F. and Wang, W.X. 2007. Fractional Differential Masks of Digital Image and Their Numerical Implementation Algorithms. *Acta Automatica Sinica*,33, 11,1128-1135.

[8] Li, M.J., Dong, Y.B. and Wang, X.L. 2014. Medical Image Edge Detection Analysis Method Based on Fractional Differential. *Advanced Materials Research*,860, 2859-2863, Trans Tech Publications.

[9] Yang, Z.Z., Zhou, J.L., Huang, M. and Yan, X.Y. 2008. Edge Detection based on Fractional Differential. *Journal Sichuan University Engineering Science Edition*,40, 1, 152-157.

[10] Rosenfeld, A. and Thurston, M. 1971. Edge and Curve Detection for Visual Scene Analysis. *IEEE Transactions on Computers*, 5, 562-569.

[11] Cafagna, D. 2007. Fractional Calculus: A mathematical Tool from the Past for Present Engineers [Past and present].*IEEE Industrial Electronics Magazine*, 1, 2,35-40.

[12] Bist, A. and Sondhi, S.2017 July, Fractional Order Differentiator Based Filter For Edge Detection of Low Contrast Underwater Images. *IJEECS*, 6, 7, 376-383. ISSN 2348-117X.

0	$\frac{v^2-v}{2}$	0
0	$-v$	0
0	1	0

(a) x-axis negative direction

0	1	0
0	$-v$	0
0	$\frac{v^2-v}{2}$	0

(b) x-axis normal direction

0	0	0
$\frac{v^2-v}{2}$	$-v$	1
0	0	0

(c) y-axis negative direction

0	0	0
1	$-v$	$\frac{v^2-v}{2}$
0	0	0

(d) y-axis normal direction

$\frac{v^2-v}{2}$	0	0
0	$-v$	0
0	0	1

(e) Right lower diagonal direction

1	0	0
0	$-v$	0
0	0	$\frac{v^2-v}{2}$

(f) Right upper diagonal direction

0	0	1
0	$-v$	0
$\frac{v^2-v}{2}$	0	0

(g) Left lower diagonal direction

0	0	$\frac{v^2-v}{2}$
0	$-v$	0
1	0	0

(h) Left upper diagonal direction

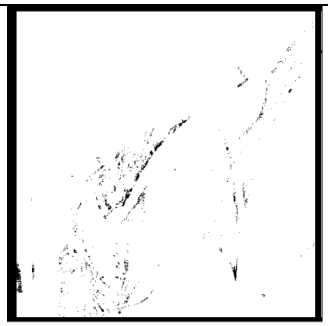
$\frac{v^2-3v}{2}$	$\frac{v^2-3v}{2}$	$\frac{v^2-3v}{2}$
$\frac{v^2-3v}{2}$	8	$\frac{v^2-3v}{2}$
$\frac{v^2-3v}{2}$	$\frac{v^2-3v}{2}$	$\frac{v^2-3v}{2}$

(i) Final Mask Template

Figure 2(a) to (i). Formulation of the Final Mask Operator

Lena Image - Original

Lena Image - Grey scale image



Lena Image - Order= 0.1

Lena Image - Order= 0.3

Lena Image - Order= 0.6

Lena Image - Order= 0.7

Lena Image - Order= 0.8

Figure 3. Resultant Images Obtained using Fractional Order Signal Processing Techniques for Different Orders

Real-time Anomaly Detection with HMOF Feature

Huihui Zhu[1], Bin Liu[2], Yan Lu[3], Weihai Li[4], Nenghai Yu[5]
School of Information Science and Technology, University of Science and Technology of China[1,2,3,4,5]
Key Laboratory of Electromagnetic Space Information, the Chinese Academy of Sciences
{zhuhui33[1], luyan17[3]}@mail.ustc.edu.cn, {flowice[2], whli[4], ynh[5]}@ustc.edu.cn

ABSTRACT
Anomaly detection is a challenging problem in intelligent video surveillance. Most existing methods are computation-consuming, which cannot satisfy the real-time requirement. In this paper, we propose a real-time anomaly detection framework with low computational complexity and high efficiency. A new feature descriptor, named Histogram of Magnitude Optical Flow (HMOF), is proposed to capture the motion of video patches. Compared with existing feature descriptors, HMOF is more sensitive to motion magnitude and more efficient to distinguish anomaly information. The HMOF features are computed for foreground patches, and are reconstructed by the auto-encoder for better clustering. Then, we use Gaussian Mixture Model (GMM) Classifiers to distinguish anomalies from normal activities in videos. Besides, visual tracking can be adopted in our algorithm for better performance. Experimental results show that our framework outperforms state-of-the-art methods, and can reliably detect anomalies in real-time.

CCS Concepts
• Computing methodologies → Artificial intelligence → Computer vision → Computer vision tasks → Scene anomaly detection

Keywords
Anomaly detection; HMOF; Auto-encoder; Real-time; Tracking.

1. INTRODUCTION
Anomaly detection and localization in intelligent video surveillance is a significant task due to the growing needs of public security. In real life, the definition of abnormalities in the video is varied. For example, a runner is seen as normal on the track and field, while it will be regarded as abnormal in the square. Therefore, it is difficult for us to use the same standard to measure all the scenes. A video event is usually considered as an anomaly if it is not very likely to occur in the video [1]. Therefore, we need the normal monitor video of the scene to establish a normal model, which identifies the anomaly in the detection.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301510

In recent studies, the sparse representations of events [2] in videos have been widely explored. The proposed models in [2] achieve favorable performance in global abnormal events (GAE), however they often fail in the local abnormal events (LAE). In order to solve this problem, some methods have proposed trajectory-based anomaly detection methods, such as particle trajectories [3], tracking trajectories [4] and so on. Such methods tend to behave very well in simple sparse scenes, but their performance in complex scenes is severely degraded. Some methods divide frames of video into numbers of patches, and then can detect anomalies in LAE by analyzing the patches [5]. In addition, some methods utilize features based on the optical flow, such as Histograms of Oriented Optical Flow (HOF) [6] and Multi-scale Histogram of Optical Flow (MHOF) [1]. These algorithms do not take into account the continuity of anomalous events in temporal and spatial domain.

Sabokrou et al [7] use two feature descriptors to extract global and local features. Leyva et al [8] extract the HOF features of the foreground area to build a dictionary and considers the joint response of the models in the local spatio-temporal neighborhood. These algorithms [7, 8] can meet the real-time requirements with high detection speeds. However, compared with other state-of-the-art methods, there are still some gaps in detection performance.

With the development of neural networks [10], anomaly detection algorithms based on deep learning develop rapidly. For example, Chong et al [11] extract spatial information from the encoder and extract temporal information using the LSTM network. DeepCascade [12] proposes a cubic-patch based method, characterized by a cascade of classifiers, which makes use of an advanced feature-learning approach.

Although there are many advantages in using deep learning to solve the problem of anomaly detection, some disadvantages exist indeed. Specifically, CNN-based method requires a large amount of training data while in anomaly detection the capacity of training set is often limited, thus it's hard to get perfect performance.

In this paper, we propose a new feature descriptor called Histogram of Magnitude Optical Flow (HMOF), which is more efficient to describe motion information. We use the foreground extraction algorithm to extract the video foreground patches, so that only the foreground patches will be processed, which is more efficient. Next, the features are fed into the auto-encoder network to be reconstructed and then classified by the Gaussian Mixture Model (GMM) Classifiers. Finally, we adopt visual tracking to capture abnormal patches for better performance.

The main contributions of our work are as follows:

(1) We present a new feature descriptor named HMOF for anomaly detection. Compared with existing feature descriptors, HMOF is more sensitive to motion magnitude and more efficient to distinguish anomaly information;

(2) To make use of the continuity of temporal and spatial for anomalous behavior, we use the tracking algorithm to track the abnormal patches, which effectively improves the performance of anomaly detection;

(3) Unlike most existing anomaly detection methods which fail to

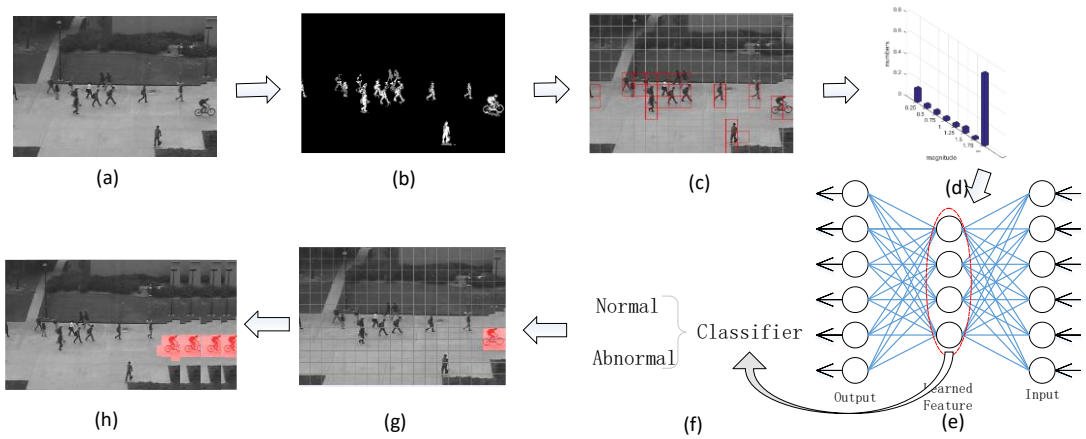


Figure. 1. Framework of our proposed method. (a) Input frames. (b) Extracted foreground. (c) Foreground patches. (d) HMOF features. (e) Auto-Encoder. (f) GMM Classifier. (g) Detected anomalies. (h) Tracking module.

balance performance and speed, the proposed framework can achieve a real-time speed and also outperform state-of-the-art methods, which can meet the demand of efficiency for video surveillance systems. In addition, we can also get better performance at the cost of sacrificing efficiency.

The rest of the paper is organized as follows. The proposed method is introduced in Section 2. Section 3 presents the experimental results, comparisons and analysis on UMN and UCSD datasets. Finally, Section 4 concludes the work.

2. THE PROPOSED METHOD

In this section, we illustrate the proposed algorithm in detail. Firstly, we obtain foregrounds patches with KNN matting. Secondly, HMOF features are extracted from foreground patches. Based on the features, we use the auto-encoder network to get the deep features, which will be fed into the GMM Classifiers. In the end, visual tracking is used to capture abnormal patches. The Framework of our method is shown in Fig.1.

2.1 Foreground Detection

We split each frame into a number of non-overlapping patches. In order to reduce the number of processing patches, we use foreground detection algorithm to extract foreground area. As a matter of fact, background patches are eliminated, which can speed up the test phase.

The problem of foreground segmentation is treated as matting, and the matting model is expressed as follows:

$$I=\alpha \mathrm{F}+(\mathrm{I}-\alpha)\mathrm{B} \qquad (1)$$

where I, F, B is color, foreground color and background color of a pixel in image respectively, α is a parameter of segmentation to indicate the former background weight. Here we use the well-known KNN matting algorithm [13] to extract the foreground. The extracted foreground is shown in Fig.1 (b). Then we calculate the foreground value of each patch by adding the intensity of every pixel. If the value exceeds a threshold, we regard it as a foreground patch. The extracted foreground patches are shown in Fig.1 (c).

2.2 HMOF Descriptor

Optical flow is good for describing the motion, and HOF is widely used as a motion descriptor. To extract HOF features, the amplitude weighting statistics of the optical flow is calculated in different directions of the optical flow, and the histogram of the optical flow direction information is obtained. However, the HOF features mainly consider the information of the optical flow direction, with less consideration of the optical flow amplitude information. MHOF [1] is based on the HOF, taking into account the optical flow amplitude information, by setting the relevant threshold information for different amplitude range of different optical flow to carry out statistics.

While MHOF uses the amplitude characteristics of the optical flow, it still mainly considers the direction of the optical flow information, which means that MHOF is less sensitive to the motion magnitude. Furthermore, the amplitude threshold is usually an experience parameter. Generally, abnormal behaviors are more sensitive to the amplitude characteristics rather than directional characteristics of the optical flow, such as running, bicycles, cars, skate, etc. The speed of these behaviors are faster than the speed of normal behaviors. To some extent, the directional characteristics of the optical flow is a kind of interference. To reach a better performance, we propose a new motion feature called HMOF based on the amplitude characteristics of the optical flow, which can detect abnormal objects effectively.

The procedure of HMOF feature extraction is as follows. Firstly, we need to calculate the threshold δ of HMOF. We sort the amplitudes of optical flow in normal patches of the whole training set in ascending order. Since there inevitably exists some noise when calculating optical flow, we discard the top ρ (the ρ value is set as 5% empirically) of the optical flow and set δ as the maximum amplitude of the remaining optical flow. Then we divide the amplitude of the optical flow into n bins. The range of i-th bin is $[(\mathrm{i}-1)/\mathrm{n}\times\delta, \mathrm{i}/\mathrm{n}\times\delta]$. In order to accommodate all the flow optical during the test phase, the range of the last bin is set to $[(\mathrm{n}-1)/\mathrm{n}\times\delta, +\infty]$. After that, we use the normalized histogram to keep the scale invariance of the HMOF features.

Fig.2 shows the feature maps of HOF, MHOF and HMOF, from left to right respectively. Among them, the pedestrian and the tree are normal, and the bicycle and the car are regarded as abnormal. It can be seen that the HMOF is more prominent than the HOF and MHOF, and the characteristic distribution is more obvious. The feature distribution of the normal region is more biased towards the low amplitude side, while the abnormal area characteristic distribution is more biased towards the high side, which is conducive to distinguish between abnormalities. Subsequent experiments show that HMOF features perform better than the HOF and MHOF features.

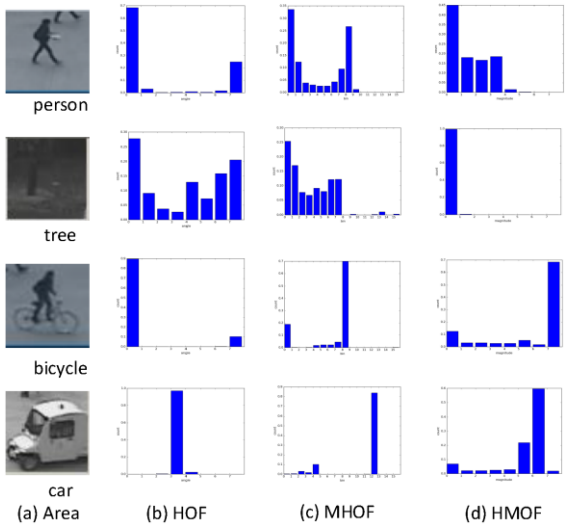


Figure. 2. The feature maps of HOF, MHOF and HMOF.

2.3 Auto-encoder

Auto-encoder, consisting of encoder and decoder, is widely used in computer vision area such as video tracking and anomaly detection. Encoder aims to project the input data into the feature space which is constructed by the hidden units, then the input is reconstructed by the decoder [14]. Auto-encoder is usually trained by minimizing the reconstruction error between the input and output. For supervised machine learning, the hidden layer of auto-encoders can be used for feature transformation.

In our case, the space expanded by the hidden units is called the feature space H. In the training phase, we use HMOF features of the training set to train the parameters of the auto-encoder. In the testing phase, we project the HMOF features of input data into feature space H. It can be seen that the features distribution of the normal and abnormal samples are quite different in the feature space H, which can be easily distinguished by the subsequent classifier.

2.4 Anomaly Classifier

The GMM is a weighed sum of multivariate Gaussian probability densities given by:

$$\mathrm{P}(\mathrm{x}\mid\theta)=\sum_{k=1}^{K}\lambda_k\,(\mathrm{x}\mid\mu_k,\Sigma_k) \tag{2}$$

where $\theta=\{\lambda_1,\cdots,\lambda_k,\mu_1,\cdots,\mu_k,\Sigma_1,\cdots,\Sigma_k\}$ denote the number of Gaussian components and λ_k is the weight of the k-th Gaussian model. μ_k and Σ_k are the mean and covariance matrix respectively. (·) denotes the multivariate Gaussian distribution. The parameters can be estimated by using the maximum-likelihood (ML) estimation. With the GMM method, we can adaptively adjust the decision surface for classification, which can better distinguish anomalies from normal activities in videos.

At first, we use all features of the training set to train the GMM Classifiers, then the trained classifier is used to test the features of the testing set. Each feature will get a score after passing the classifier. The score will be regarded as the anomaly score. If it is below a threshold α, it will be judged as abnormal (we plot Receiver Operating Characteristic (ROC) curves by using different value of α to evaluate our method):

$$Patch(\mathbf{x})=\begin{cases}Normal & score(\mathbf{x})>\alpha\\ Abnormal & otherwise\end{cases} \tag{3}$$

where X is the feature fed into the classifiers, and $score(\cdot)$ is the score given by the GMM Classifiers.

However, individual patches may be misjudged due to noise. In order to filter out these patches, we propose a method to smooth anomaly detection results spatially. For each patch judged as abnormal by GMM Classifiers, if all the eight adjacent patches around it are judged as normal, then this patch is considered to be misjudged. Furthermore, a target may be composed of several patches due to the limitation of patch-based method. Thus, we judge a frame as abnormal if the number of abnormal patches exceeds a threshold β, so as to enhance the robustness. Otherwise, these patches are more likely to be misjudged and we will drop them out from the abnormal candidates.

2.5 Tracking Module

Due to challenges such as occlusion, illumination variation, low resolution and so on, the detected abnormal behavior is often discrete in temporal and spatial domain, but the abnormal behavior under the surveillance video is often continuous. For example, there is a little boy riding a bicycle on the commercial plaza. This kind of behavior is an abnormal behavior, which is different from walking pedestrians. However, due to factors such as noise and occlusion, the abnormal behavior we detected may be intermittent, which presents discrete states in temporal and spatial.

In this paper, we use tracking algorithm to solve these problems. After each patch is determined to be abnormal or not by GMM classifiers, we use the well-known Kernel Correlation Filter [15] (KCF) to track abnormal patches to further improve the performance of our anomaly detection algorithm.

Since adding a tracking module will slow down the speed, adding this module or not depends on the system requirements. If it is a real-time monitoring system, tracking module is not necessary; if it is offline analysis system, which performance is more significant than time efficiency, we can add tracking module to get better performance.

3. EXPERIMENTS

We use two measures to evaluate the results: the frame-level and the pixel-level. For the frame-level, if one pixel is detected as an anomaly, the whole frame is considered as an anomaly. For the pixel-level, a frame is deemed to be correctly classified if at least 40% of the pixels are correctly classified [16]. Due to page limitations, we give the representative experimental results on two datasets: the UMN dataset for GAE detection and the UCSD Ped2 dataset for LAE detection. Also, two criteria are used to evaluate the ROC curves: area under curve (AUC) and equal error rate (EER). Higher AUC and lower EER indicate a better performance. The details are shown below.

Table 1. Comparison of AUC on the UMN dataset.

Method	AUC	Real-time	Method	AUC	Real-time
ZH [17]	99.3%		Leyva[8]	88.3%	√
DeepCa-scade[12]	99.6%		Sabok-rou [7]	99.6%	√
Ours-HOF	97.7%	√	Ours-tracking	**99.8%**	
Ours-MHOF	98.1%	√	Ours	**99.7%**	√

Table 2. EER comparisons on frame-level and pixel-level on UCSD Ped2 dataset; we only list first author in this table.

Method	Frame-level	Pixel-level	Real-time	Method	Frame-level	Pixel-level	Real-time
MDT [16]	24%	54%		Dan Xua [6]	20%	42%	
MPCCA [18]	30%	-		DeeCamp [12]	8.2%	19%	
Hasan [19]	21.7%	-		Sabokrou [7]	19%	24%	√
Ryota [20]	13.9%	15.9%		Biswas [21]	29.6%	42.3%	√
Tan Xiao [22]	10%	17%		Leyva [8]	19.2%	36.6%	√
Ying Zhao [23]	22%	33%		Cewu Lu [24]	22.3%	49.8%	√
Ours-HOF	21.1%	34.9%	√	Ours-tracking	**5.3%**	**11.9%**	
Ours-MHOF	19.8%	31.7%	√	Ours	**7.2%**	**14.8%**	√

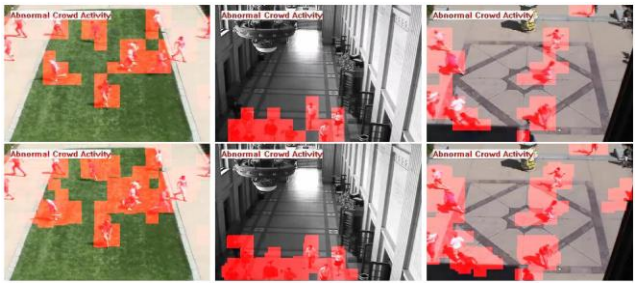


Figure. 3. Examples of anomaly detection on UMN dataset by using Ours (top row) and Ours-tracking (bottom row).

3.1 Detection of GAE on UMN Dataset

The UMN dataset has three different scenes with a resolution of 320×240. In each scene, a group of people are walking in an area, and suddenly all people run away, which is considered to be abnormal. This dataset has no pixel-level ground truth, so we use AUC of the frame-level to evaluate our method.

We set the patch at a size of 20×20, and the amplitude of the optical flow is divided into 8 bins. The threshold θ of HMOF is set to 1.04 by calculation while the number K of Gaussian components is set to 3 and the β used to adjudge abnormal frame is set to 3. Some image results are shown in Fig.3. Table 1 shows the comparison of our method with state-of-the-art methods including Sabokrou (CVPR2015) [7], Leyva (TIP2017) [8], DeepCascade (TIP2017) [12] and so on, which demonstrates that our method has a nearly perfect performance in terms of GAE detection.

3.2 Detection of LAE on UCSD Ped2 Dataset

The UCSD Ped2 dataset has 16 training and 12 testing video clips, and the number of frames of each clip varies. The videos consist of walking pedestrians paralleling to the camera plane, which are recorded with a static camera at 10 fps.

We set the patch at a size of 20×20, and the amplitude of the optical flow is divided into 8 bins. The threshold θ of HMOF is set to 2.4 by calculation while the number K of Gaussian components is set to 3 and the β is set to 3. Some image results are shown in Fig.4, which indicates that our method can detect abnormal behaviors more accurately.

Our method is compared with state-of-the-art methods including Sabokrou (CVPR2015) [7], Hasan (CVPR2016) [19], Ryota (ICCV2017) [20], Leyva (TIP2017) [8], DeepCascade (TIP2017) [12] and so on. EER comparisons on the frame-level and the pixel- level is shown in Table 2. From Table 2, we can see that if the HMOF features are replaced by HOF or MHOF in the proposed method, the performance is much worse, which indicates the effectiveness of the HMOF features.

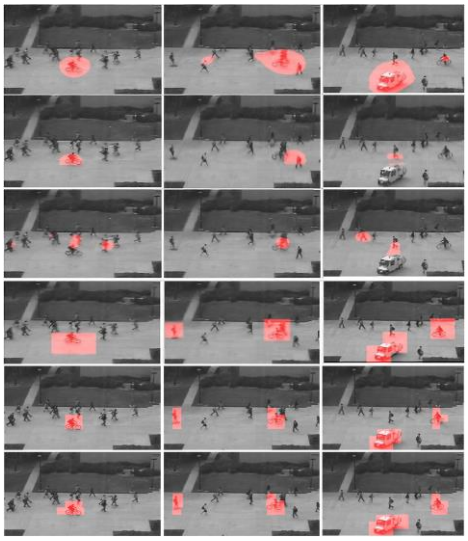


Figure. 4. Examples of anomaly detection on UCSD Ped2 dataset. Detection results from top to bottom are from Spatial MDT [16], MPPCA [18], Social Force [25], Sabokrou [7], Ours and Ours-tracking.

We can also see from Table 2 that, compared with other real-time methods, the proposed algorithm has greatly enhanced the performance, with the frame-level EER decreased from 19% to 7.2%, and the pixel-level EER from 24% to 14.8%. When compared with other state-of-the-art methods, our method also performs the best with the lowest EER in terms of both frame-level and pix-level. Besides, if we use tracking module at the expense of speed, performance has been further improved, with the frame-level EER decreasing from 7.2% to 5.3%, and the pixel-level EER from 14.8% to 11.9%.

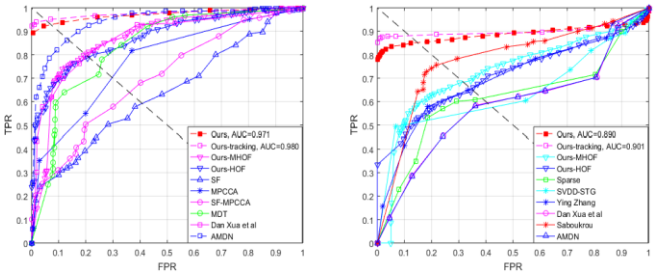


Figure. 5. ROC comparison with state-of-the-art methods. Left: Frame-level. Right: Pixel-level.

Fig.5 shows the frame-level and pixel-level ROC curves on the UCSD Ped2 dataset. The AUC of our proposed method is 97.1% on frame-level (and 89.0% on pixel-level) which, to our best knowledge, is the highest up to now. If add tracking module, performance has been further improved, with the frame-level AUC increased from 97.1% to 98.0%, and the pixel-level AUC from 89.0% to 90.1%.

4. RUUING-TIME ANALYSIS

The experiments are conducted on a regular PC with Intel-i7-7700 CPU (3.6 GHz) and 8 GB RAM, and our method is computationally efficient with the total time for detecting an anomaly in a frame on UCSD Ped dataset being 0.048 seconds per frame, which indicates that it is among the few real-time algorithms in anomaly detection community. If tracking module is added, the speed will drop to 0.25 seconds per frame. Therefore, the tracking module improves the performance at the expense of speed, adding or not adding this module depends on the system requirements.

5. CONCLUSION

In this paper, we present a novel anomaly detection method. A new feature descriptor named HMOF is proposed. Compared with other feature descriptors, HMOF is more sensitive to motion magnitude, and efficient to represent anomaly information. The frame-level EER of our proposed algorithm in UCSD ped2 dataset achieves 7.2% while the pixel-level EER is 14.8%, both of which are state-of-the-art. Moreover, it can run at 20.83fps, which meet real time demand of monitoring system. Besides, tracking module can also be adopted in our method for better performance, which shows the great potential of our method in offline analysis system.

6. REFERENCES

[1] Yang Cong, Junsong Yuan, and Ji Liu, “Sparse recon struction cost for abnormal event detection,” in CVPR, 2011 IEEE Conference on. IEEE, 2011, pp. 3449–3456.

[2] Yang Cong, Junsong Yuan, and Yandong Tang, “Video anomaly search in crowded scenes via spatio-temporal motion context,” IEEE Transactions on Information Forensics and Security, vol. 8, no. 10, pp. 1590–1599.

[3] Shandong Wu, Brian E Moore, and Mubarak Shah, “Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes,” in CVPR, 2010 IEEE Conference on. IEEE, 2010, pp. 2054–2060.

[4] Zhang Jing, Gao Wei, Liu Anan, Gao Zan, Su Yuting, and Zhang Zhe, “Modeling approach of the video semantic events based on motion trajectories,” Electronic Measurement Technology, vol. 9, pp. 008, 2013.

[5] Vikas Reddy, Conrad Sanderson, and Brian C Lovell, “Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture,” in CVPRW, 2011 IEEE Computer Society Conference on. IEEE, 2011, pp. 55–61.

[6] Dan Xu, Rui Song, Xinyu Wu, Nannan Li, Wei Feng, and Huihuan Qian, “Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts,” Neurocomputing, vol. 143, pp. 144–152.

[7] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette, “Real-time anomaly detection and localization in crowded scenes,” in CVPRW, 2015, pp.56–62.

[8] Roberto Leyva, Victor Sanchez, and ChangTsun Li, “Video anomaly detection with compact feature sets for online performance,” IEEE Transactions on Image Processing, 2017.

[9] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu, “Learning deep event models for crowd anomaly detection,” Neurocomputing, vol. 219, pp. 548–556, 2017.

[10] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” Computer Science, 2014.

[11] Yong Shean Chong and Haur Tay Yong, “Abnormal event detection in videos using spatiotemporal autoencoder,” pp. 189–196, 2017.

[12] M Sabokrou, M Fayyaz, M Fathy, and R Klette, “Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes.” IEEE Transactions on Image Processing, pp. 1992–2004, 2017.

[13] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang, “Knn matting,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 9, pp. 2175–2188, 2013.

[14] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, “Adversarial autoen-coders,” arXiv preprint arXiv: 1511.05644, 2015.

[15] Joa˜o F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, “High-speed tracking with kernelized correlation filters,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 583–596, 2015.

[16] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, “Anomaly detection in crowded scenes,” in CVPR, 2010 IEEE Conference on. IEEE, pp. 1975–1981.

[17] Yang Liu, Yibo Li, and Xiaofei Ji, “Abnormal event de-tection in nature settings,” International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 7, no. 4, pp. 115–126, 2014.

[18] Jaechul Kim and K Grauman, “Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates,” in CVPR, 2009, pp. 2921–2928.

[19] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roychowdhury, and Larry S. Davis, “Learning temporal regularity in video sequences,” in Computer Vision and Pattern Recognition, 2016, pp. 733–742.

[20] Ryota Hinami, Tao Mei, and Shin, “Joint detection and recounting of abnormal events by learning deep generic knowledge,” 2017.

[21] Sovan Biswas and R. Venkatesh Babu, “Real time anomaly detection in h.264 compressed videos,” in NCVPRIPG, 2014, pp. 1–4.

[22] Tan Xiao, Chao Zhang, and Hongbin Zha, “Learning to detect anomalies in surveillance video,” IEEE Signal Processing Letters, vol. 22, no. 9, pp. 1477–1481, 2015.

[23] Ying Zhang, Huchuan Lu, Lihe Zhang, and Xiang Ruan, “Combining motion and appearance cues for anomaly detection,” Pattern Recognition, vol. 51, pp. 443–452, 2016.

[24] Cewu Lu, Jianping Shi, and Jiaya Jia, “Abnormal event detection at 150 fps in matlab,” in ICCV, 2014, pp. 2720–2727.

[25] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In CVPR, 2009, pages 935-942

Cloud Detection for DSCOVR EPIC Data

Qing Guo
Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences
No 9 South Dengzhuang Road
Haidian District, Beijing, China,100094
guoqing@radi.ac.cn

Yanhong Zhao
Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences
No 9 South Dengzhuang Road
Haidian District, Beijing, China,100094

An Li
Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences
No 9 South Dengzhuang Road
Haidian District, Beijing, China,100094

ABSTRACT
The cloud covers most of the Earth's space and plays an important role in the Earth's energy balance. Moreover, the cloud is one of the most vital and active factors in the weather and climate. In addition, the cloud usually covers ground information, which causes many problems and difficulties in the processing of image registration and fusion. So cloud detection is very significant and necessary. DSCOVR (deep space climate observatory) satellite was launched in 2015. The new EPIC (earth polychromatic imaging camera) data from it has some special characteristics, such as the hemisphere scale and the wide range of band spectrum (from ultraviolet bands, visible bands to infrared bands). Hence, we propose a new cloud detection method for EPIC data in the way of normalized difference cloud index (NDCI). In our method, first, we analyze the different reflection characteristics of different bands, especially the new ultraviolet bands, and select appropriate bands to detect clouds. Combined with the applications of EPIC data bands, 340nm, 388nm, 680nm and 780nm are identified as the main research bands. Secondly, we analyze the reflection characteristics of clouds including thin clouds and residual clouds. Based on the above two aspects, we define the cloud index (CI) to detect clouds, which effectively reduces the influence of underlying surface on the cloud detection results. In order to verify the effectiveness of the proposed cloud detection method, other three cloud detection methods are compared, including the visible light cloud detection method, SVM cloud detection method and traditional NDCI cloud detection method. The experimental results show that our proposed method effectively detects thin clouds and residual clouds that are not detected by other methods, even in winter and in summer.

CCS Concepts
Theory of computation-->Design and analysis of algorithms

Keywords
DSCOVR EPIC; cloud detection; hemisphere scale; cloud index; cloud amount; cloud distribution.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301520

1. INTRODUCTION
The cloud covers most of the Earth's space and plays an important role in the Earth's water cycle, the energy balance of the Earth and the radiation transmission. At the same time, the cloud is one of the most vital and active factors in the weather and climate. In addition, the cloud usually covers ground information, which causes many problems and difficulties in the processing of image registration and fusion. So cloud detection is very significant and necessary. The accuracy of cloud detection affects the subsequent remote sensing applications, so improving the accuracy of cloud detection can greatly promote the practical applications.

As early as 1983, cloud detection technology was an important part of the world climate research program. Various cloud detection methods have been proposed. Most methods use the threshold to detect clouds based on the spectral characteristics of clouds. The essence of the visible light cloud detection method is to use the different reflection characteristics of clouds in different bands. More specifically, the reflectivity of the cloud in the visible light band is high, then, the threshold method is applied to separate the cloud and non-cloud. This kind of method [1-3] is simple, effective and has been widely used in cloud detection. However, only thick clouds and some thin clouds are detected in this method, the accuracy is not enough high and some missed clouds are existed. In addition, some ground objects with high reflectivity are wrongly considered as clouds.

The cloud index method generates the index between different reflectivity of different bands in the ratio way to realize the cloud detection adaptively to different satellites. The normalized difference cloud index (NDCI) method [4] is the typical one, which also uses the characteristic of different bands with different reflection features.

Support vector machine (SVM) cloud detection method is also widely used due to the advantages of few sample size, low structural risk and non-linearity [5]. However, the manually selected training samples are required in SVM, which is affected subjectively.

The DSCOVR satellite is launched by NASA and NOAA on 11, February 2015, which moves around the sun-earth Lagrange L1 point [6]. At this point, the DSCOVR satellite remains relatively stationary due to the gravitational pull of the sun and the earth. The EPIC (earth polychromatic imaging camera) is a multispectral camera carried on the DSCOVR satellite, which can see the entire earth hemisphere illuminated by the sun. Hence, the EPIC data are hemispherical scale data, which can observe the earth surface to the greatest extent possible and help the prediction and dynamic monitoring of global cloud distribution.

Therefore, this paper designs the optimal band combination to detect clouds in the index way based on the characteristics of spectral reflection and EPIC data, then, calculates the cloud amount and the cloud distribution. The SVM and the visible light cloud detection methods are compared in order to verify the effectiveness of the proposed cloud detection method. It is shown that our cloud detection method effectively detects thin clouds and residual clouds. Especially in the winter, the proposed method still effectively distinguishes clouds and snow.

2. PRINCIPLE AND METHOD

2.1 EPIC Brief Introduction

The EPIC images span 10 narrow spectral bands across ultraviolet, visible to near-infrared regions. The spatial resolution is 20km [7]. The EPIC L1 data includes L1A and L1B data. The L1A data is the raw data captured by the EPIC camera. For the L1B data, 10 bands have the same latitude and longitude, which are after rotated and rectified by L1A data [8]. In this paper, the L1B data is used as the experimental data.

The band information of EPIC data is shown in Table 1. The advantages of EPIC data for cloud detection are as follows:

(1) The cloud has different reflection characteristics in different spectral bands. The EPIC data has a wide range of wavelengths, covering violet, visible and infrared bands, which is beneficial to cloud detection.

(2) The time resolution is from 1 hour to 2 hours. Hence at least 12 scene images can be obtained in one day and the images cover almost the entire earth. Moreover, the exposure time of EPIC sensor is very short, about 60 ms to 100 ms, which makes the captured image clear.

(3) The DSCOVR satellite is relatively stable relative to the sun and the earth under the sun-earth Lagrange L1 point.

(4) The EPIC image is for the entire hemisphere surface of the earth illuminated by the sun, which effectively display the cloud distribution and change trends of cloud in the hemisphere.

Table 1. EPIC band information.

Band name (nm)	Band type	Wavelength (nm)	Full width (nm)	Usage
317	Ultraviolet	317.5±0.1	1±0.2	ozone
325	Ultraviolet	325±0.1	2±0.2	ozone
340	Ultraviolet	340±0.3	3±0.6	ozone, aerosol, cloud
388	Ultraviolet	388±0.3	3±0.6	aerosol, cloud
443	Visible	443±1	3±0.6	aerosol
551	Visible	551±1	3±0.6	aerosol, vegetation
680	Visible	680±0.2	2±0.4	aerosol,vegetation,cloud
688	Visible	687.8±0.2	0.8±0.2	cloud height
764	Infrared	764±0.2	1±0.2	cloud height, aerosol
780	Infrared	779.5±0.3	2±0.4	vegetation, cloud

2.2 Band Selection

Since the reflectivity is determined by the physical properties of various ground objects, different ground objects have different reflectivity in different bands. Moreover, the reflectivity of the cloud is higher than that of other ground objects, which is the main basis for detecting clouds. In addition, there are significant differences in the reflectivity of different clouds in each band. Generally speaking, the reflectivity of the thick cloud is highest, followed by the thin cloud, and the residual cloud has the lowest reflectivity.

EPIC images have unique violet bands, namely 317 nm, 325 nm, 340 nm and 388 nm. Violet can be used to distinguish clouds, snow and other features. The 317 nm and 325 nm bands are commonly used to study ozone in the atmosphere. The 340 nm and 388 nm bands are used for cloud research. Hence, the 340 nm and 388 nm bands are selected for study in the violet range.

The visible light band (680 nm, 551 nm, 443 nm) also shows strong reflectivity for cloud, snow and water. Herein, the 443 nm is commonly used for aerosol research, the 551 nm is often used to study aerosols and vegetation, and the 680 nm band can be used for cloud research. So, the 680 nm band is used as the research band in our paper.

The oxygen absorption band (688 nm, 764 nm) is a unique band of EPIC, which is often used to invert the cloud height. Therefore, the 688 nm and 764 nm bands are not considered as our research band in this paper.

The near-infrared band (780 nm) is commonly used to distinguish cloud and snow. In the near-infrared band, the reflectivity of the cloud is higher, and the reflectivity of snow is lower. So, 780 nm is our research band.

Except for the visible light bands, other bands alone cannot be used for cloud detection. Therefore, it is necessary to combine two or more bands for cloud detection. For example, the dust index [9] is used to distinguish between cloud and dust. The normalized snow index or snow index [10] is used to distinguish between cloud and snow.

The traditional threshold cloud detection method can detect most of thick clouds, but it is not sensitive to the detection of thin clouds and residual clouds. In addition, since the snow has a high reflectivity, the traditional threshold cloud detection method may determine snow or other ground objects with higher reflectivity as clouds, which causes misjudgment of the cloud detection. However, the cloud index method effectively reduces the impact of other ground objects on cloud detection. Hence, in this paper, the optimal band combination of 340 nm, 388 nm, 680 nm, and 780 nm is studied in the cloud index way to realize the cloud detection for EPIC hemisphere data.

2.3 Cloud Detection Method for EPIC Data

The normalized cloud index (NDCI) is based on the definition of the normalized vegetation index (NDVI). The definition of the NDVI is as follows:

Herein, ρ_{NIR} and ρ_{RED} are the reflectivity of the near-infrared and the red bands, respectively.

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}} \quad (1)$$

Approximately, the NDCI is defined as the ratio of the reflectivity difference between the near-infrared band and the visible band to the reflectivity sum between two bands [11].

The cloud index method uses the bands sensitive to clouds in the reflectivity ratio way to detect clouds. Based on selected research bands in 2.2, we design the optimal band combination for cloud

detection. Since the wavelength difference of 340 nm and 388 nm is not large, two band combinations CI_{340} and CI_{388} are proposed:

$$CI_{340} = \frac{\rho_{680} - \rho_{340}}{\rho_{780}}$$
$$CI_{388} = \frac{\rho_{680} - \rho_{388}}{\rho_{780}} \quad (2)$$

It is noted that the EPIC L1B data is the data obtained after the corrections of earth rotation and the positional offset of the spacecraft based on the original data.

However, the band combination requires the reflectivity value corresponding to each band. So, it is necessary to multiply the calibration factor provided in Table 2 to convert the level 1 data into the reflectivity. The accuracy of the calibration factor is about 1-3%. [12].

Table 2. The calibration factors for each EPIC band.

Band	Calibration factor
317	1.216E-04
325	1.111E-04
340	1.975E-05
388	2.685E-05
443	8.340E-06
551	6.660E-06
680	9.300E-06
688	2.020E-05
764	2.360E-05
780	1.435E-05

3. EXPERIMENTS AND ANALYSIS

Due to the high reflectivity of snow, the cloud detection in winter is mainly due to snow interference, which may cause large differences. In addition, there are more residual clouds and thin clouds in summer. In order to verify the effectiveness of the cloud index method proposed in this paper, the experimental data with two representative times are chosen, including the defined winter (January 3, 2017) and summer (July 3, 2017) data.

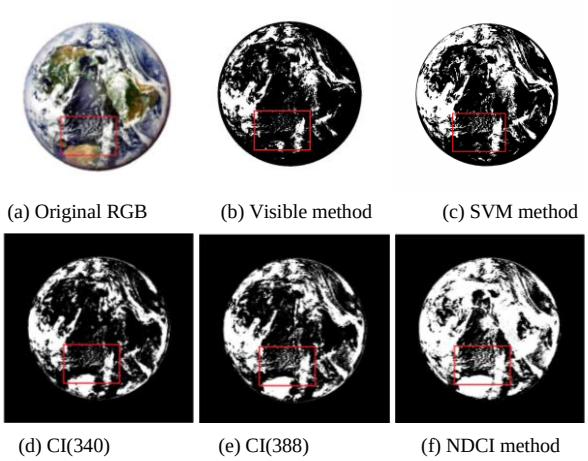
The RGB band combination true color image contains almost all colors that human vision can perceive. This RGB color system is one of the most widely used color systems. Therefore, the original RGB color EPIC images are used as the visual references for cloud detection.

In order to further compare the effectiveness of cloud detection methods, the visible method, SVM method and NDCI method are also compared. From both the cloud distribution and the cloud amount aspects, the experimental results are analyzed. In the cloud distribution binary images, cloud pixels are marked as 1, non-clouds are marked 0.

The visible method adopts the red band (680nm) to detect clouds. For the SVM method, it is found that the detection accuracy is improved to use the HSV (hue-saturation-saturation) color space image than the RGB color space image [13]. Hence, in our experiments, the RGB color space image is first transformed to the HSV color space image, then, the SVM method is used to detect clouds based on the HSV color image.

For the NDCI method, the near-infrared band and the visible band are required. In our experiment, the visible 443 nm band is adopted. Comparing to the visible 551 nm and 680 nm bands, the 443 nm band is used to detect clouds, which can reduce the misjudgment of land area as clouds. However, even using the 443nm band, in the cloud distribution result, it still detects the land area with high reflectivity as clouds leading to misjudgment and wrong judgment. So, in the cloud amount comparison, the result of the NDCI method is not considered.

The cloud detection results for the defined summer and winter of five different methods are shown in Figure 1 and Figure 2, respectively. The cloud amount comparison results are given in Table 3 and Table 4, respectively.



(a) Original RGB (b) Visible method (c) SVM method

(d) CI(340) (e) CI(388) (f) NDCI method

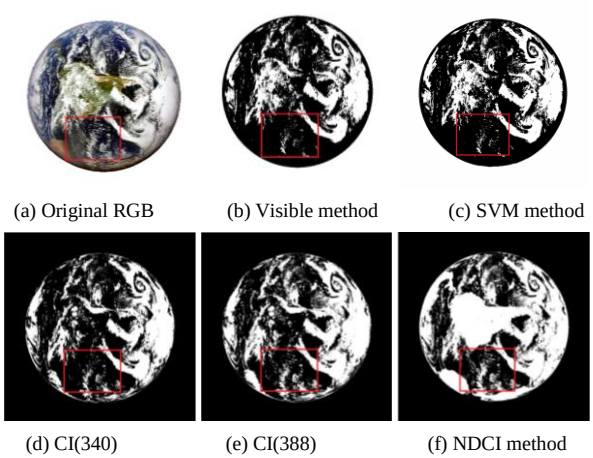
Figure 1. The cloud detection comparison of different methods for the defined summer: (a) the original RGB image; (b) the visible method; (c) the SVM method; (d) the proposed CI (340) method; (e) the proposed CI (388) method; (f) the NDCI method.

Table 3. Comparison of cloud amount for the defined summer.

Methods	Visible method	Cloud index mehotd		SVM method
	680	CI(340)	CI(388)	With HSV transform
Clouds(Pixel)	392094	500204	583821	602325
Non-clouds(Pixel)	1468446	1468446	1359145	1251833
Cloud amount (%)	21.07	26.90	31.40	32.49

From both the cloud distribution and the cloud amount, the experimental results are analyzed. Referencing to the original RGB image and the cloud distribution image, the distributions of the detected clouds in the visible method, the SVM method and the proposed method are basically consistent with that in the original RGB, in which the human vision can distinguish clouds. In particular, the cloud distribution in the proposed method is the closest to the actual result. The visible method is insensitive to thin clouds and residual clouds, which causes the missed detection and does not detect the clouds in the red frame. The proposed method effectively reduces the influence of the underlying surface through the ratio calculation between bands. The proposed method detects the thin clouds and residual clouds, which are not detected by the visible method. The detection accuracy of the SVM method is unstable due to the random selection of samples. In the Figure 1(c), the SVM method detects some thin clouds and residual clouds. But in the Figure 2(c), it does not detect the thin clouds and residual clouds in the red frame. In the NDCI method,

the land area with high reflectivity is wrongly determined as the clouds. Therefore, in the proposed CI method, the cloud distribution has the best visual effect.



(a) Original RGB (b) Visible method (c) SVM method

(d) CI(340) (e) CI(388) (f) NDCI method

Figure 2. The cloud detection comparison of different methods for the defined winter: (a) the original RGB image; (b) the visible method; (c) the SVM method; (d) the proposed CI (340) method; (e) the proposed CI (388) method; (f) the NDCI method.

Table 4. Comparison of cloud amount for the defined winter.

Methods	Visible method	Cloud index method		SVM method
	680	CI(340)	CI(388)	With HSV transform
Clouds (Pixel)	614714	708952	772273	627917
Non-clouds (Pixel)	1394066	1296949	1233628	1375800
Cloud amount (%)	30.60	35.34	38.50	31.34

In the aspect of cloud amount shown in Tables 3-4, for the data in July, the cloud amounts of the visible method, the proposed CI(340) and CI (388), and the SVM method are 21.07%, 26.90%, 31.40% and 32.49%, respectively. The cloud amount in the visible method is the lowest. The biggest difference of other methods is 5.59%. It is evident that the visible method misses a large number of thin clouds and residual clouds. For the data in January, the cloud amounts are 30.60%, 35.34%, 38.50% and 31.34%, respectively. The biggest difference of the later three methods is 7.16%. The cloud amount difference between the two proposed CI methods is small. In the January data, there is more snow than that in July. So, it is evident that the snow has a great impact on cloud detection, especially in the visible method. However, the use of CI form effectively reduces the misjudgment of snow in cloud detection and improves the detection accuracy.

4. CONCLUSION

The new EPIC data from the DSCOVR satellite have the hemisphere scale, the wide range of band spectrum and some new bands. Due to these aspects, the CI cloud detection is defined to effectively reduces the influence of underlying surface on the cloud detection results. The method is analyzed from two aspects: the cloud amount and the cloud distribution. In order to verify the effectiveness of the proposed cloud detection method, other three cloud detection methods are compared, including the visible light method, SVM method and traditional NDCI method. The EPIC data corresponding to the defined summer (July 3, 2017.) and winter (January 3, 2017.) are used to conduct experiments. The results of cloud distribution show that our proposed method effectively detects thin clouds and residual clouds that are not detected by other methods, even in winter and in summer. The cloud distribution obtained by our proposed method is most consistent with the cloud distribution in the original EPIC image with the combination of RGB true color.

5. ACKNOWLEDGMENTS

Our thanks to the Atmospheric Sciences Data Center of NASA Langley Research Center for allowing us to download the EPIC data. The downloading address for EPIC L1 data is: https://eosweb.larc.nasa.gov/project/dscovr/dscovr_epic_l1b_2.

This work was supported in part by the National Natural Science Foundation of China under Grants 61771470 and 41590853; in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences under Grant QYZDY-SSWDQC026; and in part by the Youth Innovation Promotion Association, Chinese Academy of Sciences under Grant 2014054.

6. REFERENCES

[1] Wen T., He M. Y., Zhao Z. L., Yao Z. G., Han Z. H., Kuang S. S. 2016. Research on cloud detection method based on GMS-5 satellite data. *Infrared*, 37(2):29-35. DOI= http://10.3969/j.issn.1672-8785.2016.02.005.

[2] Wang Q., Sun L., Wei J., Zhou X. Y., Chen T. T., Shu M. Y. 2018. Improvement of dynamic threshold cloud detection method and high resolution satellite application. *Acta Optica Sinica*, 38(10):1028002-1.

[3] Deng S., Li G., Zhang H. 2017. Objective determination scheme of threshold in high-spectral-resolution infrared cloud detection. Meteorological *Monthly*, 43(2): 213 -220. DOI= http://10.7519/j.issn.1000-0526.2017.02.009.

[4] Wen X. F., Dong X. Y., Liu L. M. 2009. Cloud index method for cloud detection. *Geomatics and Information Science of Wuhan University*, 34(7):838-841.

[5] Ishida H., Oishi Y., Morita K. 2018. Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions. *Remote Sensing of Environment*, 205: 390-407. DOI=http:// 10.1016/j.rse.2017.11.003.

[6] Yang Y., Marshak A., Mao J., Lyapustin A., Herman J. 2013. A method of retrieving cloud top height and cloud geometrical thickness with Oxygen A and B bands for the Deep Space Climate Observatory (DSCOVR) mission: Radiative Transfer Simulations. *Journal of Quantitative Spectroscopy & Radiative Transfer*, 122(6):141-149. DOI=http://dx.doi.org/10.1016/j.jqsrt.2012.09.017.

[7] Yang B., Knyazikhin Y., M õttus M., Rautiainen M., Stenberg P., Yan, L., Chen, C., Yan, K., Choi, S., Park, T., Myneni, R.B. 2017. Estimation of leaf area index and its sunlit portion from DSCOVR EPIC data: Theoretical basis. *Remote* Sensing *Environment*, 198, 69–84. DOI=http://dx.doi.org/10.1016/j.rse.2017.05.033.

[8] EPIC Geolocation and Color Imagery Algorithm Revision 5, https://eosweb.larc.nasa.gov/project/dscovr/DSCOVR_EPIC _Geolocation_V02.pdf

[9] Hai Q. S., Bao Y. H., Alateng T. Y., Bao G., Guo L. B. 2009. New method to identify sand and dust storm by using remote sensing technique---With Inner Mongolia autonomous region as example, *Journal of Infrared and Milimeter Waves*,

28(2):129–132. DOI=http://10.3321/j.issn:1001-9014.2009.02.012.

[10] Lin, J., Feng X., Xiao P., Li H., Wang J. Li Y. 2012.Comparison of snow indexes in estimating snow cover fraction in a mountainous area in northwestern China, *IEEE Geosci. Remote Sens. Lett.* 9(4), 725–729. DOI=http://10.1109/LGRS.2011.2179634.

[11] Marshark A., Knyazikhin Y., Davis A., Wiscombe W., Pilewskie P. 2000. Cloud–vegetation interaction: Use of normalized difference cloud index for estimation of cloud optical thickness. *Geophysical Research Letters*, 27(12), 1695–1698. DOI=http://10.1029/1999GL010993.

[12] EPIC calibration factors v02, https://eosweb.larc.nasa.gov/project/dscovr/DSCOVR_EPIC_Calibration_Factors_V02.pdf

[13] Li W., Li D. R. 2011. The cloud detection study of MODIS based on HSV color space. *Journal of Image and Graphics*, 16(9):1696-1701.

A FPN-Based Framework for Vehicle Detection in Aerial Images

Yinjuan Gu
Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
School of Communication and Information Engineering, Shanghai University, Shanghai, China
guyinjuan@sina.cn

Bin Wang
Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
School of Communication and Information Engineering, Shanghai University, Shanghai, China
brantley.wang@brantley.wang

Bin Xu
Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China
School of Communication and Information Engineering, Shanghai University, Shanghai, China
15621092258@163.com

ABSTRACT
Vehicle detection in aerial images plays an important role in many aspects, including traffic surveillance, urban planning, parking lot analysis, etc. However, due to the influence of low resolution, complex background and rotating objects, vehicle detection in aerial images still has limited progress. In this paper, we propose a specialized framework for vehicle detection in aerial images. In this framework, Feature Pyramid Network and Faster RCNN are combined to utilize both low-resolution, semantically strong features and high resolution, semantically weak features to achieve a better detection rate of small objects. On this basis, focal loss function is adopted to reduce the imbalance between easy and hard samples. The experiments show that the performance of proposed framework is better than other state-of-the-art frameworks.

CCS Concepts
• Computing methodologies → Object detection

Keywords
Vehicle detection, feature pyramid network, aerial images

1. INTRODUCTION
Vehicle detection in aerial images is a branch of object detection, which is the central topics in the computer vision and photogrammetry literature. It is important for various applications, e.g., traffic management, urban planning, parking lot utilization, etc. However, in contrast to vehicle detection in ground view images, vehicle detection in aerial images is still confronted with many challenges and difficulties, such as low spatial resolution (i.e., small numbers of pixel on target), complex backgrounds and monotonic appearance.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301531

Object detection has always been a fundamental problem of computer vision. With the rapid development of deep learning techniques, in particular, the convolutional neural network, CNN features have been gradually adopted to improve the accuracy of object detection instead of traditional hand-crafted features. Since then, the series of methods based on region convolutional neural network ([1] [2] [3] [4]) have pushed forward the progress of object detection. However, in the specific domain of vehicle detection, like detection in aerial images, the methods mentioned above are not applicable ([12]). In other words, a more pertinent network structure should be employed for vehicle detection in aerial images. There are two main reasons why it is the case. First of all, the vehicles in aerial images are usually quite small (a car might be only 30×12 pixels) ([21]). Second, the complex background of man-made objects, which appear visually similar to the cars is also an important factor. Hence, we focus on a specific framework, FPN (Feature Pyramid Network) ([5]), which is applied to detect various scales of objects, including objects in aerial images.

The contributions of this paper are presented as follows: (1) A specialized framework which combines Feature Pyramid Network with Faster RCNN is proposed and applied to vehicle detection in aerial images. (2) An annotation method is designed to be more suitable for the detection task. (3) A more applicable loss function is adopted in the proposed framework.

The reminder of this paper is organized as follows: In Section 2, a brief review of vehicle detection in aerial images is presented. In Section 3, we make a detailed introduction about the proposed framework. The VEDAI dataset and annotation method are detailed in Section 4, with experiments and results elaborated in this part. In Section 5, we discuss the concluding remarks and future work.

2. RELATED WORK
Prior to the development of deep learning, sliding window detectors ([8]) were the state-of-the-art approach to object detection. Sliding window methods utilize both a specific hand-crafted feature representation such as histogram of gradients (HOG) and a classifier such as a support vector machine (SVM) to independently binary classify all sub-windows of an image as belonging to an object or background([9] [10]). Even through their methods report good results, hand-crafted features are insufficient to separate vehicles from complex background.

Most of the existing methods for vehicle detection in aerial images are mainly based on the deep learning ([11] [12] [13]). In

general, object detection methods can be roughly classified into three main steps: candidate region proposal, feature extraction and classification. The methods which consist of these three steps are well-known as two-stage methods, such as the series of methods based on region convolutional neural network (RCNN). Sakla et al. modified Faster RCNN parameters to make the network adjusted to handle the small, challenging objects that are presented in VEDAI dataset ([14]). Tang et al. used enhanced RPNs that utilize a shallow fine feature map and split large target images into small tiles ([24]). Koga et al. proposed using hard example mining (HEM) in the training process of a convolutional neural network and got good results ([23]). In contrast, the methods which do not need an additional operation for region proposal, such as YOLO ([15]) and SSD ([16]), are one-stage methods. However, the one-stage methods do not perform well in small objects. This demerit limits their application for vehicle detection in aerial images.

3. METHODOLOGY

The proposed framework is shown in Figure 1. It consists of two main parts: Faster RCNN and Feature Pyramid Network (FPN). The backbone structure we choose is ResNet ([8]). As we can see in the figure, FPN-based RPN and FPN-based ROI Pooling are the most important two parts. FPN-based RPN is designed to generate region proposals and in this part, we use focal loss function for classification. FPN-based ROI Pooling is used to extract features.

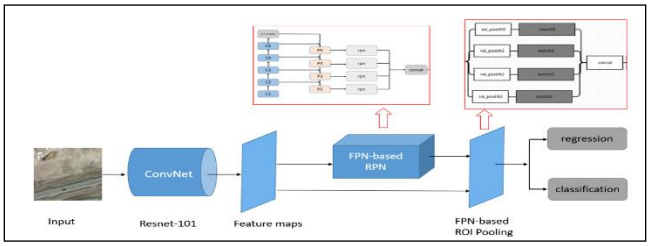


Figure 1 The overview of the proposed framework

3.1 Faster RCNN

Faster RCNN is one of the state-of-the-art detection frameworks whose simplicity and robustness have attracted a wide range of researchers since 2015 (Sakla et al., [14]). It has been proven that Faster RCNN achieves the good detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets ([3]). An overview of Faster RCNN is illustrated in Figure 2. It is an optimized version of the paper ([2]). Instead of selective search ([2]), Faster RCNN utilize a Region Proposal Network (RPN) for generating high-quality region proposals. And then the learned region proposals are fed upstream into the Fast RCNN detection network.

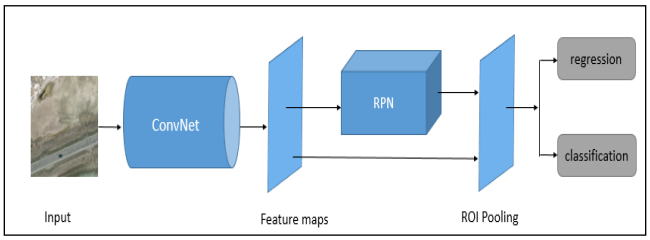


Figure 2 The overview of Faster RCNN

3.1.1 RPN

Region Proposal Network is a fully convolutional network that simultaneously predicts object bounding boxes and *objectness* scores at each position. The input of RPN is an n×n spatial window of the input convolutional feature map (usually n=3). The outputs of this convolutional layer's sliding window are mapped to a low-dimensional feature with ReLU following. Finally, the low-dimensional features are fed into two fully-connected layers (regression layer and classification layer). The detailed explanations about RPN can be found in [3].

3.1.2 Resnet-101

While Faster RCNN can utilize the convolutional layers from any CNN architecture, we choose the Resnet-101 model ([8]). The reason is that increasing the depth of the neural network can improve the performance of our network. Many articles ([17] [18]) have revealed that the network depth is of crucial importance and some nontrivial visual detection tasks have also greatly benefited from very deep models. Resnet-101 has 5 groups of convolutional layers. The first group contains one 7×7 kernels and the structure of the remaining four groups is illustrated in table 1.

Table 1 The structure of ResNet-101

Layer Name	Type	101-layer	Output_Size
Conv1	C1	7×7, 64, stride2	112×112
Conv2_x	C2	3×3 max pool, stride2	56×56
		[1×1, 64; 3×3, 64; 1×1, 256] ×3	
Conv3_x	C3	[1×1, 128; 3×3, 128; 1×1, 512] ×4	28×28
Conv4_x	C4	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	14×14
Conv5_x	C5	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	7×7
		average pool, 1000-d fc, softmax	1×1

3.2 Feature Pyramid Network

Traditional methods ([19]) use an image pyramid to build a feature pyramid and features are computed on each of the image scales independently, which is computation intensive. Most of the recent detection methods ([2][3][20]) use only single scale features to improve detecting speed so that the accuracy of detecting small objects is declined. In this paper, we implement a special framework, Feature Pyramid Network (FPN), which can make a good balance between detecting speed and accuracy.

The task of semantic segmentation has proved that features from the lower layers retain more detailed information ([22]). Compared to the objects in ground view images, objects in aerial images hold much less pixels and smaller portion in the original images ([21]). Therefore, when detecting vehicles in aerial images, features from the shallow layers can be fused with those from deeper layers to improve the detection accuracy. The author of FPN explains why FPN improves features for small objects. On the one hand, we need high-resolution feature maps to obtain more contextual information; on the other hand, as we can see in Figure 3, more detailed information should be provided to determine whether the car exists and where it is ([5]).

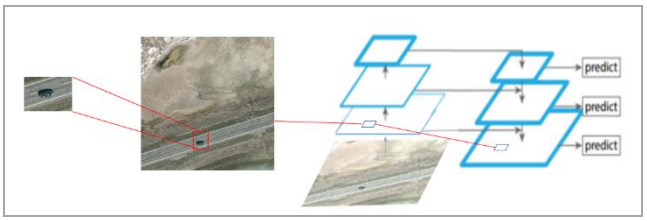


Figure 3 Illustration of vehicle detection in aerial images

Feature pyramid network can combine low-resolution, semantically strong features with high resolution, semantically weak features via a top-down pathway and lateral connections. As is shown in Figure 4, the construction of the pyramid involves a bottom-up pathway, a top-down pathway, and lateral connections. The bottom-up pathway is the feed-forward computation of the backbone ConvNet (In this paper, the backbone ConvNet is Resnet-101). For Resnet-101 we use the feature activations output by each stage's last residual block and denote the output of these last residual blocks as {C1, C2, C3, C4, C5} for conv1, conv2, conv3, conv4, conv5 outputs (see Table 1). We do not include C1 into our pyramid network because of its large memory footprint. The top-down pathway hallucinates higher resolution features by upsampling spatially coarser, but semantically stronger feature maps from higher pyramid levels. These features are enhanced with features from bottom-up pathway via lateral connections. The final set of feature maps is {P2, P3, P4, P5}, corresponding to {C2, C3, C4, C5}.

In this paper, we apply FPN to Faster RCNN in order to obtain good results in aerial image detection. Faster RCNN is the combination of Fast RCNN and RPN. Thus, we adopt FPN in RPN for bounding box proposal generation and in Fast RCNN for object detection. The diagrammatic sketch of FPN-based Faster RCNN are shown in Figure 5.

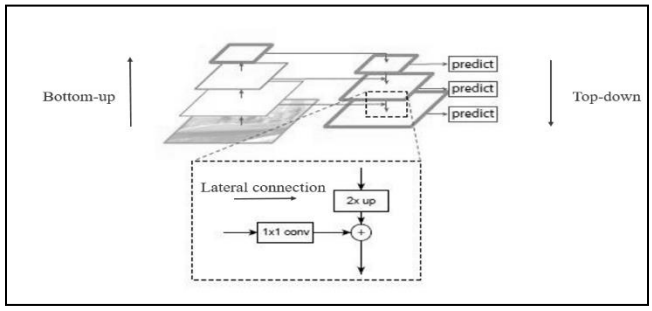


Figure 4 The framework of FPN

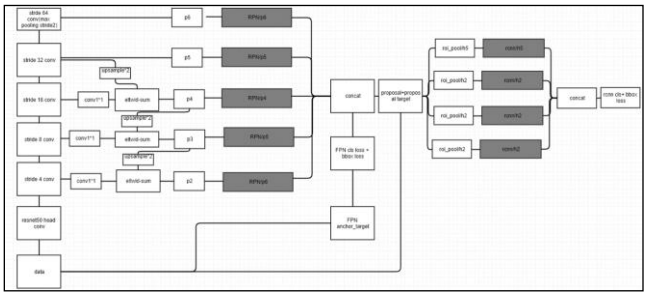


Figure 5 FPN-based Faster RCNN

3.3 Loss Function

Cross entropy loss function is the most popular loss function used for vehicle detection. It describes the distance between two probability distributions, and the more similar the two are when the cross entropy is smaller. It can improve the imbalance between positive and negative samples to a certain extent. However, as for the distinction between easy and hard examples, it doesn't perform well. It is hard to select the negative samples which are very similar as the vehicles and if we can't solve this problem, the result of our detection may be influenced as well. Thanks to the appearance of focal loss function ([6]), we change the loss function in the region proposal stage.

The original loss function of region proposal stage is formally defined as:

$$L_{\text{rpn}} = \frac{1}{N_{cls}} \sum L_{cls}\left(p_i, \mathrm{p}_i^*\right) + \lambda \frac{1}{N_{reg}} \sum \mathrm{p}_i^* L_{smo}$$

$$L_{smo}(i) = \begin{cases} 0.5i^2 & if\ |\mathrm{i}| < 1 \\ |i| - 0.5 & otherwise \end{cases}$$

$$L_{cls} = -\log\left(p_t\right), \quad \mathrm{p}_t = \begin{cases} p & \text{if } \mathrm{y} = 1 \\ 1-p & otherwise \end{cases}, \ \mathrm{y} \in \{-1, +1\}$$

where L_{cls} is the conventional cross entropy and L_{smo} is the smooth loss function. N_{cls} and N_{reg} denote the total number of samples and the total number of positive samples respectively. We change the L_{cls} and it is defined as:

$$L_{cls} = -\alpha(1-p_t)^{\gamma} \log\left(p_t\right)$$

where α=0.25 , γ=2 in this paper. Focal loss function is an optimization of the CE function. It is obviously that the easier the sample is, the larger the p_t and the smaller the loss. This formula L_{cls} distinguishes the easy /hard samples.

4. EXPERIMENTS

4.1 VEDAI Dataset and Annotation

All experiments are performed using VEDAI dataset, which consists of more than 1200 images that come in two different resolutions 512×512 and 1024×1024. The imagery is available in three visible color channels and one near infrared channel. All the images in the paper are conducted on the infrared version of the 512×512 images for the reason that 512×512 images can provide smaller, more challenging targets. While the VEDAI dataset contains 9 classes of objects, we focus our attention on small vehicles, namely the car, pickup, and van classes. A total of 2440 instances across these three classes are presented in 995 images of the VEDAI dataset. We randomly choose 80% of the images for training and the remaining 20% for testing. In addition, the original annotation of VEDAI dataset includes the centroid, orientation, and coordinates of four corners of each instance. In this paper, we change the way of annotation in the following manner: retain the centroids and generate 20×20 pixel square bounding box around the centroids for the 512×512 resolution imagery. As for the performance evaluation, we use the standard precision and recall statistic to compute the average precision metric. F1-score is also used in our evaluation system as an

important reference index. The definitions of these metrics are formally described as:

$$\text{Recall Rate(R)} = \frac{TP}{TP + FN}$$

$$\textit{Precision Rate(P)} = \frac{TP}{TP + FP}$$

$$F1\_\text{score} = \frac{2 \times R \times P}{R + P}$$

Where, TP, FN, FP denote the true positive, false negative and false positive respectively

4.2 Results on VEDAI Dataset

We first report our results on VEDAI using the original architecture. All experiments are carried out by modifying the Python re-implementation of the Faster RCNN and FPN repository using Tensorflow framework. We use the initial learning rate of 0.001 and a momentum value of 0.9 and train the models for a total of 70k iterations. Besides, we use the pretrained model ResNet-101 for our network. The other detailed parameters and initial result are presented in Table 2.

Table 2 Initial parameter of the proposed network

parameter	value
BASE_ANCHOR_SIZE_LIST	[32,64,128,256,512]
ANCHOR_SCALES	[1]
ANCHOR_RATIOS	[0.5,1,2]
SCALE_FACTORS	[10,10,5,5]

Table 3. shows us the detecting result of our framework and the comparison between the result using cross entropy loss function and that using focal loss function.

Table 3 The comparison between CE and FL results

	RECALL	PRECISION	AP	F1
Result(CE)	0.806	0.907	0.75	0.853
Result(FL)	0.864	0.893	0.77	0.878

Table 4 The comparisons with related works

METHOD	mAP
DPM Razakarivony and Jurie[2015]	60.5
SVM+HOG31 Razakarivony and Jurie[2015]	55.4
SVM+LBP Razakarivony and Jurie[2015]	51.7
SVM+LTP Razakarivony and Jurie[2015]	60.4
SVM+HOG31+LBP Razakarivony and Jurie[2015]	61.3
SVM Fusion AED(HOG) Razakarivony and Jurie[2014]	69.6
Faster-RCNN Sakla W, Konjevod G[2017]	70.9
Ours(CE)	**75.4**
Ours(FL)	**77.2**

Our method reaches a mAP of 0.772, which is a good improvement (6.3%) from the previous state of the art. As a comparison, the recent work of [14] reports a mAP of 0.7093 in the same class using Faster RCNN. The Table 3 lists all the published results on VEDAI and ours. We can see that in terms of Average Precision the advantages of our method.

Figure 6 shows us the detecting results directly. The right column is the ground truth of original picture which is annotated in the required form. The left column is the result of our detecting algorithm. It is obviously that the proposed detection framework performs well on VEDAI dataset.

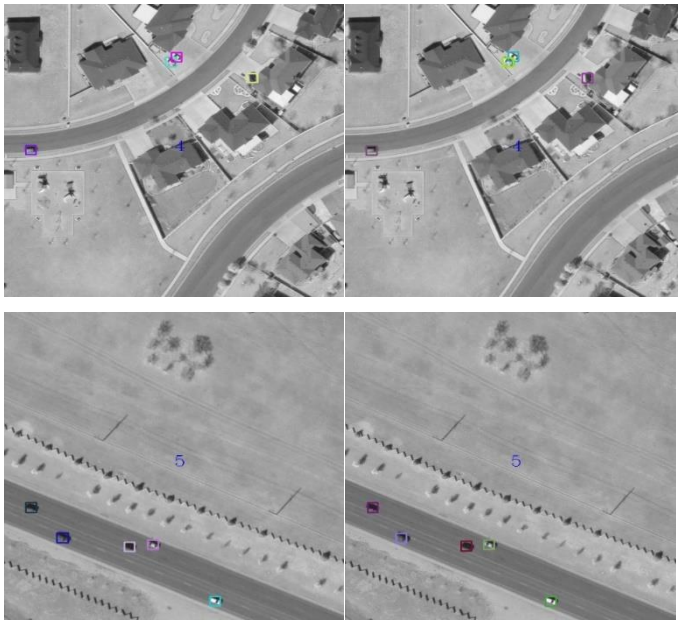


Figure 6 The detecting result using proposed network

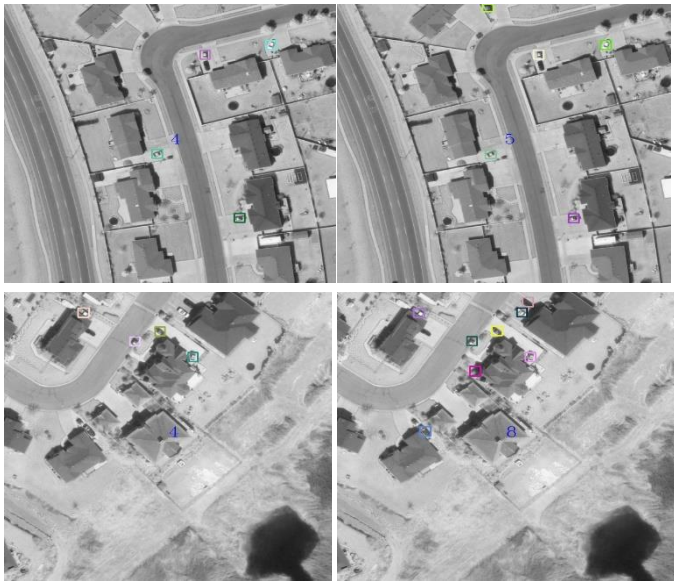


Figure 7 CE result (left) and FL result (right)

As for the result using modified network (using focal loss function instead of cross entropy loss function), Figure 7 gives us a detailed comparison. The left column uses CE loss function to detect vehicles, while the right column uses focal loss function. Compared to the algorithm using CE loss function, the algorithm using focal loss function can detect more 'true positive' in the same figure.

5. CONCLUSION

In this paper, a specialized framework which combines FPN with Faster-RCNN is proposed to detect small objects. The backbone of the network is changed (using Resnet-101 instead of VGG16 for vehicle detection in aerial images). The model is trained and tested in the VEDAI dataset. Besides, in order to improve the imbalance between easy and hard examples, the focal loss function is used instead of conventional cross entropy loss function. To make a comparison, we applied the same dataset to Faster-RCNN, and the experimental results show that our method outperforms the state-of-the-art algorithm. For future work, we will pay more attention on improve the accuracy of vehicle detection in aerial images and we think that changing the annotation method and taking orientation into consideration may be good ideas.

This work was supported by the National Natural Science Foundation of China (grant number: 61601280) and Ministry of Education of China (grant number: P201606).

6. REFERENCES

[1] Girshick R, Donahue J, Darrell T, et al. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation* [J]. 2013:580-587.

[2] Girshick R. *Fast R-CNN* [J]. Computer Science, 2015.

[3] Ren S, He K, Girshick R, et al. *Faster R-CNN: towards real-time object detection with region proposal networks*[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015:91-99.

[4] He K, Gkioxari G, Dollar P, et al. *Mask R-CNN*. [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP (99):1-1.

[5] Lin T Y, Dollar P, Girshick R, et al. *Feature Pyramid Networks for Object Detection*[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017:936-944.

[6] Lin T Y, Goyal P, Girshick R, et al. *Focal Loss for Dense Object Detection* [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP (99):2999-3007.

[7] Razakarivony S, Jurie F. *Vehicle detection in aerial imagery: A small target detection benchmark* ☆ [J]. Journal of Visual Communication & Image Representation, 2015, 34:187-203.

[8] He K, Zhang X, Ren S, et al. *Deep Residual Learning for Image Recognition* [J]. 2015:770-778.

[9] Gleason, Joshua, et al. "*Vehicle detection from aerial imagery*." IEEE International Conference on Robotics and Automation IEEE, 2011:2065-2070.

[10] Razakarivony S, Jurie F. *Discriminative Auto encoders for Small Targets Detection*[C]// International Conference on Pattern Recognition. IEEE, 2014:3528-3533.

[11] Chen X, Xiang S, Liu C L, et al. *Vehicle Detection in Satellite Images by Parallel Deep Convolutional Neural Networks*[C]// Iapr Asian Conference on Pattern Recognition. IEEE Computer Society, 2013:181-185.

[12] Jean Ogier Du Terrail, Frédéric Jurie. *ON THE USE OF DEEP NEURAL NETWORKS FOR THE DETECTION OF SMALL VEHICLES IN ORTHO-IMAGES*. IEEE International Conference on Image Processing, Sep 2017, Beijing, China

[13] Tang T, Zhou S, Deng Z, et al. *Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining*[J]. Sensors, 2017, 17(2):336.

[14] Sakla W, Konjevod G, Mundhenk T N. *Deep Multi-modal Vehicle Detection in Aerial ISR Imagery*[C]// Applications of Computer Vision. IEEE, 2017:916-923

[15] Redmon J, Divvala S, Girshick R, et al. *You Only Look Once: Unified, Real-Time Object Detection* [J]. 2015:779-788.

[16] Liu W, Anguelov D, Erhan D, et al. SSD: *Single Shot MultiBox Detector*[C]// European Conference on Computer Vision. Springer International Publishing, 2016:2

[17] K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition.* In ICLR, 2015.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan,V. Vanhoucke, and A. Rabinovich. *Going deeper with convolutions*.In CVPR, 2015

[19] E.H. Andelson and C.H. Anderson and J.R. Bergen and P.J. Burt and J.M. Ogden. "*Pyramid methods in image processing*". 1984

[20] He K, Zhang X, Ren S, et al. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1904-1916.

[21] Liu K, Mattyus G. *Fast Multiclass Vehicle Detection on Aerial Images* [J]. IEEE Geoscience & Remote Sensing Letters, 2015, 12(9):1938-1942.

[22] Shelhamer E, Long J, Darrell T. *Fully Convolutional Networks for Semantic Segmentation* [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, PP(99):1-1

[23] Koga Y, Miyazaki H, Shibasaki R. *A CNN-Based Method of Vehicle Detection from Aerial Images Using Hard Example Mining* [J]. Remote Sensing, 2018, 10(1).

[24] Tianyu Tang, Shilin Zhou, Zhipeng Deng, et al. *Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining*[J]. Sensors (Basel, Switzerland), 2017, 17(2):336.

DSBI: Double-Sided Braille Image Dataset and Algorithm Evaluation for Braille Dots Detection

Renqiang Li
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
lirenqiang@ict.ac.cn

Hong Liu*
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
hliu@ict.ac.cn

Xiangdong Wang
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
xdwang@ict.ac.cn

Yueliang Qian
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
ylqian@ict.ac.cn

ABSTRACT
Braille is an effective way for the visually impaired to learn knowledge and obtain information. Braille image recognition aims to automatically detect Braille dots in the whole Braille image. There is no available public datasets for Braille image recognition to push relevant research and evaluate algorithms. This paper constructs a large-scale Double-Sided Braille Image dataset DSBI with detailed Braille recto dots, verso dots and Braille cells annotation. To quickly annotate Braille images, an auxiliary annotation strategy is proposed, which adopts initial automatic detection of Braille dots and modifies annotation results by convenient human-computer interaction method. This labeling strategy can averagely increase label efficiency by six times for recto dots annotation in one Braille image. Braille dots detection is the core and basic step for Braille image recognition. This paper also evaluates some Braille dots detection methods on our dataset DSBI and gives the benchmark performance of recto dots detection. We have released our Braille images dataset on the GitHub website.

CCS Concepts
• Computing methodologies → Object recognition

Keywords
Braille image; dataset, auxiliary annotation, Braille dots detection, benchmark.

1. INTRODUCTION
In the world, there are about 1.3 billion people with vision impairment and 36 million people are blind according to the World Health Organization in 2018 [1]. Braille is an effective way for the visually impaired to learn knowledge, obtain information and communicate with other people.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301532

Braille document consists of Braille characters and each Braille character has a rectangular block called Braille cell, which contains six Braille dots arranged in three rows and two columns with 64 different combinations [2]. Many Braille books are double-sided in order to save pages, which may contain recto dots and verso dots in one Braille image. Braille image recognition aims to automatically detect Braille dots in the whole Braille image and recognize Braille cells to Braille characters.

Many existing Braille image recognition methods are based on image segmentation and used several designed rules to discriminate the Braille dots [3, 4]. Some work used statistical learning methods to recognize Braille images [5, 6]. However, above Braille dots detection and Braille cell recognition methods are tested on their small-scale Braille images datasets with different acquisition ways and most of them are based on single-sided Braille. The performance on more complex, various Braille books and the large-scale dataset is lack.

There is no available public datasets for Braille image recognition to push relevant research and evaluate existing methods. This paper focuses on constructing such a large-scale Braille image dataset with double-sided, which can also provide Braille dots and Braille cells annotation information. And we also implement and evaluate some Braille dots detection methods on this dataset.

There are many ways to acquire Braille images from original Braille documents, including the camera, scanner and some special devices. Schwarz et al. [7] designed a controlled lighting environment to acquire Braille images with a fixed camera and fixed light sources, which is inconvenient and difficult to realize. Zhang et al. [8] used the camera of mobile phone to capture Braille images, which will be disturbed by background illumination and bring much distortion to captured images. Antonacopoulos et al. [3] used the flat-bed scanner to obtain Braille images, which is a quite simple and convenient way and with less image distortion. So we also adopt the general flatbed scanner to capture our Braille images.

The main contributions of this paper are as follows:

(1) We construct the first public Double-Sided Braille image dataset DSBI with Braille recto dots, verso dots and Braille cells annotation. This dataset includes 114 double-sided Braille images from 6 Braille books and some ordinary printed documents for keeping diversity and complexity. It is available on the Github website: https://github.com/yeluo1994/DSBI.

(2) An auxiliary annotation strategy is proposed to quickly annotate Braille images, which adopts initial automatic detection

of Braille dots and modifies annotation results by convenient human-computer interaction method. This labeling strategy can averagely increase label efficiency by six times for recto dots annotation in one Braille image.

(3) Braille dots detection is the core and basic step for Braille image recognition. We evaluate some Braille dots detection methods on our Braille image dataset DSBI and give the benchmark performance of recto dots detection. These methods include Braille dots detection based on image segmentation and Haar+Adaboost classifier.

Our DSBI dataset is discussed in Section 2. Section 3 presents the proposed auxiliary annotation method. Experimental results analysis and conclusions are drawn in section 4 and section 5.

2. BRAILLE IMAGE DATASET

The purpose of constructing this dataset is to provide a benchmark for the current Braille recognition methods on real complex double-sided image. We will describe our construction method in details.

2.1 Braille Image Acquisition Way

As mentioned in the introduction section, there are many ways to acquire Braille images and different acquisition ways have different influence on the performance of Braille image recognition. We select flatbed scanner to obtain the double-sided Braille images, which is convenient and can provide good quality of Braille images. The scanner used in this paper is HP LaserJet Pro MFP M226dn. To reduce storing memory and remaining enough clarity, we use 200dpi resolution to capture color Braille images and store them in JPEG format. And we scan two sides of each Braille document to get recto dots and verso dots Braille information.

2.2 Braille Image Quality

Figure1 shows a local region sample of double-sided Braille image captured by our general HP scanner, which has uniform illumination. And it has good captured quality for Braille images recognition methods.

But in double-sided Braille images, some recto dots and verso dots are mixed together, which are difficult to achieve high-precision results of Braille image recognition. And to enhance diversity and complexity, our Braille images are acquired from several Braille books including different background document color, different production ways and different usage degree. Figure 2 shows some samples with above situation and some defects in our dataset, such as oil stains in Figure 2(a), paper distortion in Figure 2(b), some cracks in Figure 2(c), and abrasion Braille dots in Figure 2(d). These difference and defects are actually very common in real applications, which can also evaluate different methods with more objectivity.

2.3 Description of DSBI

Table 1 gives the detailed description of our constructed Braille image dataset DSBI. There are total 114 color double-sided Braille images from six different Braille books and six ordinary printed Braille documents.

These Braille books include reference books, such as Massage, professional textbooks, such as middle school Chinese textbook, and novels, such as Shaver Yang Fengting. To effectively evaluate Braille dots detection methods, we also add some pages containing only Braille verso dots in our dataset. And some Braille images have defects, such as oil stains from Massage book

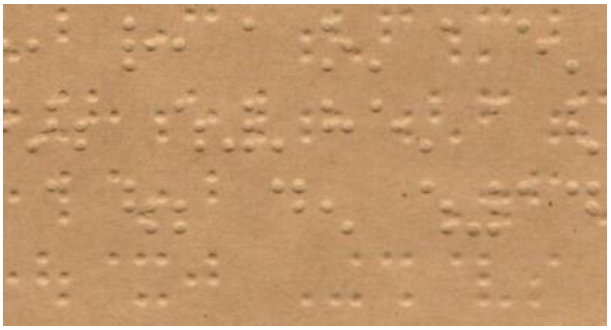
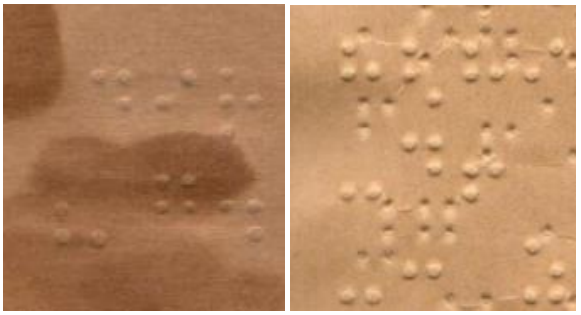
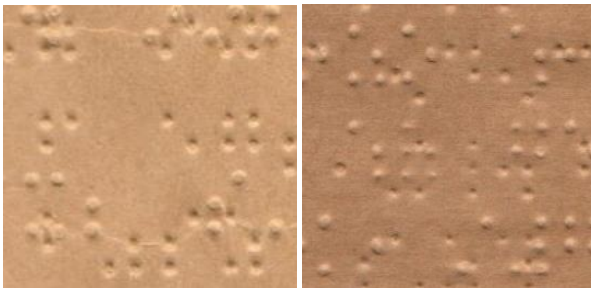


Figure. 1 A region sample of double-sided Braille image.



(a) With oil stain **(b) With paper distortion**



(c) With some cracks **(d) With abrasion Braille dots**

Figure 2. Complex samples of some local Braille images.

Table 1. Detailed description of proposed dataset DSBI.

Book name	Total pages	Verso dots pages	Image quality
1.Massage	20	1	Bad
2.Fundamentals of Massage	20	1	Normal
3.The Sec. Volume of Ninth Grade Chinese Book 1	20	1	Normal
4.The Second Volume of Ninth Grade Chinese Book 2	10	1	Normal
5.Math	32	0	Normal
6.Shaver Yang Fengting	6	0	Normal
7.Ordinary printed document	6	0	Good
Total	114	4	N/A

as Figure2 (a) shows. And the image quality from each book is also shown in Table 1.

For different Braille production ways and image capture noise, the Braille images we captured by scanner maybe exist some skewed angles. So our DSBI provides de-skewing Braille images and skewed angles for each image. And to quantitatively evaluate Braille image recognition performance, including Braille dots detection and Braille cells recognition, we provide corresponding Braille recto dots, Braille verso dots and Braille cells annotation information for each double-sided Braille image in our dataset DSBI.

3. AUXILIARY ANNOTATION METHOD

Accurate annotation of Braille dots and Braille cells for each Braille image in DSBI is important for developing recognition algorithms and evaluating performance. Our DSBI dataset contains double-sided Braille images and each image may contain about hundreds of recto dots or verso dots.

It is time-consuming to label each Braille dot and Braille cell information by a manual manner. We developed an interactive labeling tool, which will averagely cost over one hour to label only recto dots for one Braille image by the manual manner. For some Braille dots are too small to identify and some verso dots will seriously disturb the estimation of recto dots, which may bring many annotation errors. And manually labeling Braille cells information is a more complicated task. To reduce workload of labeling, this paper proposes a Braille image auxiliary annotation strategy as follows.

Firstly, we use the Haar+Adaboost and sliding windows strategy to automatically detect Braille recto dots in Braille images.

Secondly, the Braille dots may not be aligned in vertical and horizontal directions for production and capture noise. A Braille de-skewing process is used to correct Braille image. We get this skewed angle by analyzing the statistics information of row and column projections of detected Braille dots under different angles.

Thirdly, the distance between dots within a Braille cell and the distance between adjacent Braille cells are relatively fixed according to the acquisition resolution. We get the location of Braille cells based on above layout rules of Braille cells. We construct a Braille cell gird to generate preliminary Braille recto dots detection results.

Finally, based on above Braille recto dots and Braille cells location results, we propose a convenient interactive Braille annotation method using numeric keys and direction keys. Direction keys can quickly move Braille cell by Braille cell row and Braille cell column, and numeric keys can quickly modify the Braille dots information in each Braille cell to ensure the validity of annotation result. Figure 3 shows one sample of Braille image auxiliary annotation process.

For verso dots annotation, a recto dot on the front page is the verso dot on the back page for the double-sided Braille image. So we can easily obtain the verso dots annotation information by the recto dots annotation on the back page. Our Braille dataset DSBI provides the recto dots, verso dots and Braille cells location annotation information.

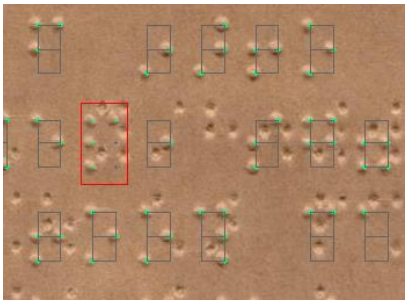


Figure 3. A sample of Braille image auxiliary annotation.

Table 2. Details of training set and test set of DSBI.

Book name	Training set pages	Test set pages
1.Massage	10	10
2.Fundamentals of Massage	0	20
3.The Second Volume of Ninth Grade Chinese Book 1	0	20
4.The Second Volume of Ninth Grade Chinese Book 2	0	20
5.Math	10	22
6.Shaver Yang Fengting	3	3
7.Ordinary printed document	3	3
Total	26	88

4. EVALUATION OF BRAILLE DOTS DETECTION METHODS

4.1 Evaluation Metrics

This paper used Precision, Recall and F1 value to evaluate the performance of Braille dots detection methods. The three indexes are as follows:

$$Precison = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{3}$$

Where TP represents the number of dots correctly identified, FN represents the number of dots misidentified as the background, and FP represents the number of background misidentified as Braille dots.

4.2 Training Set and Test Set

To debug the parameters of the segmentation method and obtain training samples for machine learning based methods, we divide our DSBI dataset into training set and test set. The total Braille images number of DSBI is 114, we select about 1/4 amount, 26 images, as training set and 3/4 amount, 88 images, as test set. Training images are from selected professional tool books, middle school textbooks, novels and ordinary printed documents. Test images are from each captured Braille book.

The descriptive information of training set and test set is shown in Table 2.

4.3 Braille Dots Detection Methods

4.3.1 Based on image segmentation

Antonacopoulos et al. [3] used a local adaptive threshold to segment gray double-sided Braille image into three parts, including highlight, shadow and background, and obtained black and white region as Figure.4 shows. They divided the region whose width exceeded a threshold and then identified vertically an adjacent pair of white-black combination as recto dot, and black-white combination as verso dot.

We implement Braille recto dots detection method based on [3] and improve the original method. To reduce capture noise by Braille document edge for segmentation, we replace the unexpected pixel values with the Braille image background pixel value. Then gray normalization is adopted to optimize image quality and reduce the influence of different Braille images background. Gray histogram is used to find the adaptive global segmentation threshold for each Braille image. Finally, we identify vertically adjacent pairs of white-black regions to detect recto dots according to the size of white region and distance to the center of black region. This method is different from [3], which simply judges a recto dot according to whether a white region exists above the black region within the expected distance. We take the matched rectangle center as the position of recto dots as Figure.5 shows.

4.3.2 Based on Haar+Adaboost

The cascade classifier using Haar feature [9], frequently used in face detection and object detection, can get fast detection performance. This paper also used Haar+Adaboost method for Braille recto dots detection.

Haar feature can be calculated quickly by the integral image technology and cascade classifier. The cascade classifier can fast reject most of negative samples in the previous classifiers, which can greatly speed up the Braille dots detection speed. Meanwhile the cascade classifier has the advantage of generalization ability to ensure the high accuracy. We used Haar extraction and training function in OpenCV to train the cascaded classifier with selected positive samples and negative samples from training set based on annotation information. The sample size and sliding windows size are all 20×20 pixels, and the sliding step is 2 to get better performance.

4.4 Results and Analysis

In order to effectively compare the Braille dots detection methods in Section 4.3, we tested on the de-skewing Braille images in our test set from DSBI dataset. We calculate the Precision, Recall, F1 value of each method on the 88 double-sided Braille images as Table 3 shows.

(a) Segmented Braille image (b) Corresponding dots

Figure 4. Segmented Braille Dots sample from [3].

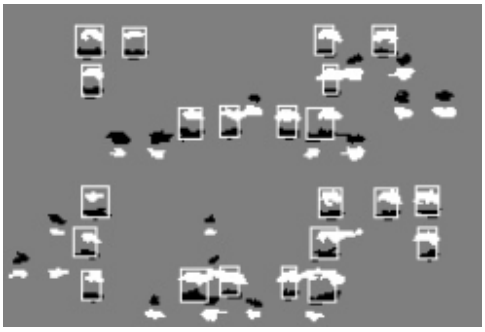


Figure 5. Results of matched recto dots by segmentation.

Table 3. Comparison of Braille recto dots detection methods

Method	Precision	Recall	F1
Based on Image segmentation	91.72%	98.11%	0.948
Based on Haar+Adaboost	97.65%	96.38%	0.970

It can be seen from the experimental results that the overall results of the two methods are relatively ideal. The Braille dots detection method based on image segmentation got 0.948 F1 value, which is the lowest among the methods. The segmentation based method will be subject to noise interference and need to set a lot of thresholds manually, which make it difficult to get better results in complex double-sided Braille images. The Braille dots detection method based on Haar+Adaboost got 0.97 F1 value, which is better than above image segmentation based method.

The Braille dots detection methods in this paper are carried out using c++ and OpenCV in the hardware environment of Intel Core I7 3.40 Ghz CPU and 16GB memory.

Above experimental results are based on Braille recto dots detection on the de-skewing double-sided Braille image. In our DSBI dataset, we also provide annotation information of verso dots and Braille cells, which can be used to test detection performance of verso dots and Braille cells. And besides de-skewing Braille images, we also provide the original double-sided Braille images and skewed angles information for each Braille image, which can be used for the de-skewing experiment.

5. CONCLUSION

This paper introduces our constructed Braille image dataset DSBI, which contains 114 double-sided Braille images from a variety of Braille books and documents. In addition, we also propose a Braille image auxiliary annotation method based on the automatic recognition of Braille images and modification of annotation results by convenient human-computer interaction method. Finally, we also implement the Braille dots methods based on image segmentation and Adaboost, and conduct experiments on our dataset DSBI to provide benchmark results. In future work, we will pay more attention to improving the performance of Braille dots detection and Braille cell location.

6. REFERENCES

[1] W. H. Organization, Visual impairment and blindness, 2018, [online] Available: http://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] Isayed S, Tahboub R. A review of optical Braille recognition[C]//Web Applications and Networking, 2015 2nd World Symposium on. IEEE, 2015: 1-6.

[3] Antonacopoulos A, Bridson D. A robust Braille recognition system [C]//International Work- shop on Document Analysis Systems. Springer Berlin Heidelberg, 2004: 533-545.

[4] S. D. Al-Shamma and S. Fathi, "Arabic braille recognition and transcription into text and voice," in Biomedical Engineering Confer- ence (CIBEC), 2010 5th Cairo International. IEEE, 2010, pp. 227–231.

[5] Yin Jia. The key technology research on paper-mediated Braille automatic recognition system [Master degree thesis]. Changchun University of Science and Technology, Changchun, 2011 (in Chinese).

[6] Li Ting. A Deep Learning Method for Braille Recognition. Computer and Modernization, 2015, 36:37-40 (in Chinese).

[7] Schwarz T, Dolp R, Stiefelhagen R. Optical Braille Recognition[C]//International Conference on Computers Helping People with Special Needs. Springer, Cham, 2018: 122-130.

[8] Zhang S, Yoshino K. A braille recognition system by the mobile phone with embedded camera[C]//Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on. IEEE, 2007: 223-223.

[9] Viola. P. and Jones. M., "Rapid object detection using a boosted cascade of simple features", In IEEE Conference on Computer Vision and Pattern Recognition 2001.

A Natural Scene Edge Detection Algorithm Based on Image Fusion

Weichao Ding
Southeast University-Monash University Joint Graduate School
Southeast University
Suzhou, China
220163499@seu.edu.cn

Zhile Yang
Shenzhen Institute of Advanced Technology
Chinese Academy of Sciences
Shenzhen, China
zl.yang@siat.ac.cn

Liangbing Feng*
Shenzhen Institute of Advanced Technology
Chinese Academy of Sciences
Shenzhen, China
lb.feng@siat.ac.cn

ABSTRACT
Convolutional neural network (CNN) has been widely used in the edge detection areas and shown competitive results. However, with the increase of receptive fields, the convolution features in CNN gradually become rough and difficult to figure out. To tackle with the problem, a novel network is proposed in this paper, making full use of the multi-scale and multi-level information of the object to perform image-to-image prediction, and combining all distinctive convolution features in a holistic manner. Further, the effect of simply connecting the feature map is enhanced by an image fusion algorithm to improve the utilization of features. The feature maps obtained by convolutions of each layer are fused through the fusion network to obtain a more detailed feature. The improved algorithm is validated in the BSDS500 dataset and the ODS F-measure has reached 0.818, which significantly exceeds the current state-of-the-art results.

CCS Concepts
• **Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems → Edge Detection**

Keywords
neural network; natural scene; edge detection; image fusion;

1. INTRODUCTION
Edge detection has been a traditional research topic in the field of computer vision. As a basic problem in the image domain, the solutions to edge detection can also provide assistance and reference for many traditional problems, such as salient object detection [1], image segmentation [2], and skeleton extraction [3].

Moreover, it also plays an important role in modern applications such as autopilot. The early edge detection algorithms mainly extract the edges of the image through physical features such as brightness, texture, and color of the image. Although these approaches which are using low- level features have made great improvement in these years [4], their limitations are obvious given the pressing demand from applications. For example, edges and boundaries are often defined to be semantically meaningful. However, it is difficult to use low-level cues to represent object-level information. Under these circumstances, gPb [5] and Structured Edges [6] are proposed using complex strategies to capture global features as much as possible.

In the past few years, convolutional neural networks have provided compelling results in advanced visual tasks such as image classification [7] or object detection [8]. Recent work also shows that CNNs can be applied to pixel-level marking tasks, including semantic segmentation or normal estimation. A suitable characteristic of these tasks is that the inherent convolutional nature of CNNs allows a simple and effective "fully convolutional" implementation. Since CNNs have the powerful ability to automatically learn advanced natural image representations, there has recently been a tendency to use convolutional networks to perform edge detection. Some well-known CNN-based methods have been applied like N^4–Fields [9], DeepContour [10], HED [11], CED [12] and RCF [13].

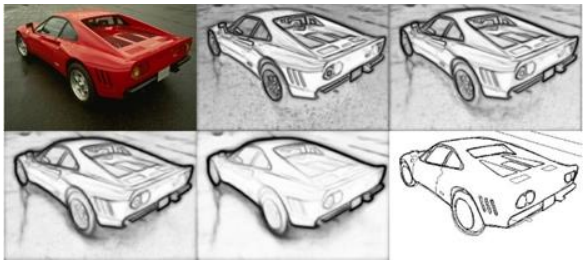


Figure 1. The original picture and the results obtained after each fusion.

In order to see the information obtained in the different convolutional layers of the edge detection, a simple network is established using vgg16 [14] to generate the side output of the middle layer, which has five phases. Many researchers have worked hard to develop deeper networks because of the richer convolutional features that are very effective for many visual tasks [15]. However, due to the disappearance/explosion gradient and insufficient training data, it is difficult to for the model converging when going deeper into the network. Unlike the previous edge detection algorithms, feature utilization rate extracted by neural networks is significantly enhanced in the proposed method. Instead of simply connecting the model structure, we use a lightweight neural network to fuse each layer's side output. This method makes the features of the image to be better used, and the edges of the resulting image are also clearer. When evaluating the proposed method on BSDS500 dataset [5], we achieve the best trade-off between effectiveness and efficiency with the ODS F-measure of 0.818, which is higher than the state-of-the-art algorithm.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00.
https://doi.org/10.1145/3301506.3301533

2. RELATED WORK

The edge detection is recognized as one of the most basic issues in computer vision [16], where researchers have struggled for nearly 50 years and have made significant progresses. Early pioneering methods focused on the use of intensity and color gradients. Roberts [16] introduced the famous Roberts operator in 1965. It uses local difference operators to find the edges in the image. The Sobel operator [17] was proposed by Sobel in 1968. It is a discrete differential operator that combines Gaussian smoothing and differential derivation to calculate a stepwise approximation of the image luminance function. An extended version of Sobel, named Canny [18], took Gaussian smoothing as a preprocessing step and used double thresholds to obtain edges. By this means, Canny is shown to be more powerful for noise. In fact, due to its remarkable efficiency, Canny is also popular in various tasks. However, these early methods seem to have poor accuracy and are therefore difficult to be adopted in the recent applications.

Later, the researchers began to construct features artificially through the low-level features of the image, and used complex learning methods to determine whether each pixel belonged to the edge of the image. [19]. Martin et al. [20] formulated Pb features that respond to changes in brightness, color, and texture, and trained a classifier to combine these features. Arbelaez et al. [??] developed Pb into gPb by using standard Normalized Cuts to combine above local cues into a globalization framework. Lim [??] proposed novel features, Sketch tokens that can be used to represent the mid-level information. Dollar et al. [??] proposed structured random forest that simultaneously learns the clustering and mapping, and directly outputs a local edge patch. However, all the above methods are based on manual functions that have limited capabilities for semantically meaningful edge detection representing high-level information.

With the rapid development of deep learning in recent years, a series of methods based on deep learning have been invented. An N^4 field combining CNNs and nearest neighbor search is proposed by Ganning et al. [9]. Shen et al. [10] divide the outline data into subclasses and fit each subclass by learning model parameters. With the HED algorithm proposed by Xie et al. [11], the end-to-end idea was applied to the edge detection domain for the first time. The edge detection algorithm has reached a new stage, and many improved algorithms based on this algorithm continue to emerge, and the accuracy of edge detection is kept to be pushed towards new heights. Our algorithm is also an improved algorithm based on this idea.

3. PROPOSED NETWORK

The edge detection network we employed can be divided into two parts including feature extraction network and feature fusion network. Such network structure increases the width of the entire network in the case of constant depth and improves the utilization of the features without losing the underlying information of the picture, due to which it is more suitable for edge detection problems.

3.1 Feature extraction network

Feature extraction network uses an improved VGG16 network structure, which has been proven in many papers to have excellent results for edge detection algorithms. Edge detection is a relatively low-level problem in computer vision compared with semantic segmentation. Therefore, many low-level features must be preserved when dealing with this problem. However, networks with deeper depths tend to lose this low-level information and pay more attention to high-level information of images. Therefore, it is appropriate to use VGG16 network as the basic network for feature extraction.

Moreover, we have made some modifications to VGG16, making it more suitable for edge detection problems. We removed the pool5 in VGG16 and the last three full-connected layers, retained

13

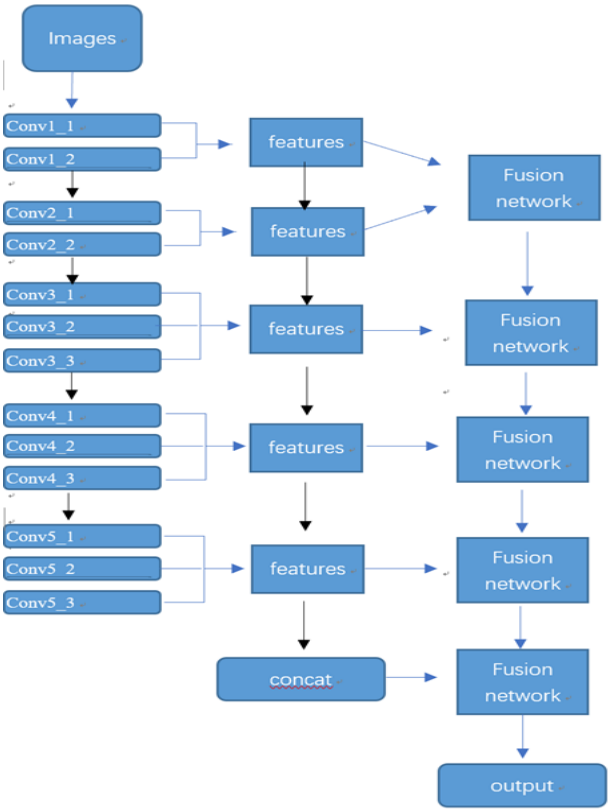


Figure 2. Our Feature extraction network architecture.

convolutional layers and 4 pooling layers. Each stage of convolution followed a 2*2 pooling layer. Previous studies have shown that the feature maps obtained by convolving each layer is used instead of using only the features obtained in the last layer of each convolution layer to effectively improve the accuracy of edge detection. This result was also verified in our experiments, so we used the feature maps obtained by convolution in each layer to sum the convolutions of the layers in each stage and then up-sampled by a 1*1 convolution. Finally, the up-sampling results of the five phases are connected to obtain the results of the sixth feature extraction phase.

3.2 Feature fusion network

The second part we designed shows obvious differences from other edge detection networks. Existing edge detection methods are usually simply connecting the features extracted by the neural network to obtain the result. However, this simple method does not make good use of the extracted features of each layer. From the perspective of

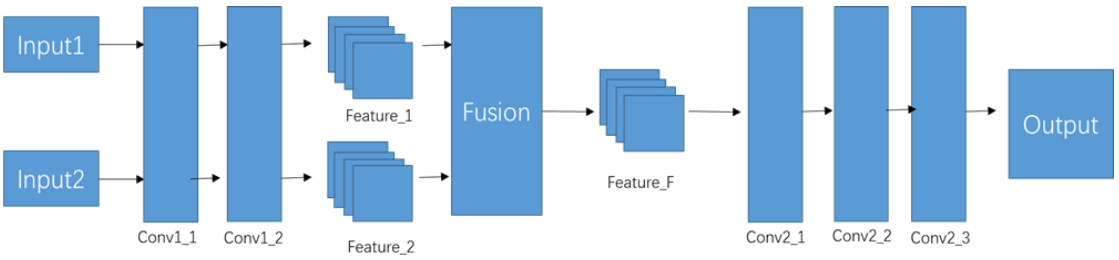


Figure 3. Our Feature fusion network architecture

HED to RCF development, improving the utilization of features has significantly increased the effectiveness of edge detection. With the continuous development of deep learning, the depth of the neural network

is gradually deepened. Because deeper networks can extract higher-level features, this trend has significantly improved high-level image problems such as semantic segmentation and object classification. However, for the low-level problems such as edge detection, deeper networks will ignore some of the underlying texture features instead. Therefore, in this paper, we try to improve the utilization of features without increasing the depth of the network.

The image fusion algorithm has been well applied to the multiple exposure fusion problem [21]. This problem is a common low-level problem in the image field. Recently, an algorithm for multiple exposure fusion using deep learning has obtained a very good performance on this issue [22]. Its network structure is simple and the computational cost is low. Inspired by this method, the image fusion algorithm is incorporated into the edge detection algorithm to improve the utilization of features. We input the feature map obtained by the fusion network layer by layer. The fusion layers C1 and C2 share the same weight, and low-level features are extracted from the input feature map. The composition layer merges the feature pairs of the input feature map into one feature. The fused features are input into the reconstruction layer to generate a fused new feature map. After layer-by-layer fusion, we obtained the final edge detection result, which is a significant improvement over the results without the image fusion algorithm.

Figure 4. Some examples on BSDS500 dataset and NYUD dataset.

4. EXPERIMENT

We use the open source framework Caffe, which is well known in this community, to implement our network. Pre-trained VGG16 models on ImageNet are used to initialize our network. The weights of the 1×1 conv layers in the fusion phase are initialized to 0.2, and the bias is initialized as 0. Stochastic Gradient Descent (SGD) 10 samples are randomly sampled in each iteration for small batches.

In evaluating our algorithm, we used two common data sets, including the most widely used BSDS500 as well as the NYUD data set. For a given edge probability map, a threshold is needed to produce an edge image. There are two major techniques to set this threshold. The first is called the best data set scale (ODS), which uses a fixed threshold for all images in the data set. The second is called the optimal image scale (OIS), which selects the optimal threshold for each image. We use both methods at the same time. Before calculating the F-measure, we also need to go through the non-maximum suppression to get a refined edge.

4.1 BSDS500 Dataset

BSDS500 is a widely used data set for edge detection. It consists of 200 trainings, 100 verifications, and 200 test images, each labeled with 4 to 9 annotators. We use the training and validation set to fine-tune and use the test set for evaluation. The parameters increase in data is the same as HED. We rotated the image to 16 different angles and cropped the largest rectangle in the rotated image; we also flipped the image at each angle, resulting in an enhanced training set 32 times larger than the unenhanced training set. During the test, we operated on the input image at its original size. Inspired by previous work, we mixed BSDS500 enhancement data with flipped Pascal VOC context datasets into training data. When evaluated, standard non-maximal suppression (NMS) is applied to thin detection edges. We compare our approach with some non-deep learning algorithms (including Canny, gPb-UCM, SE, and OEF [23], and some recent deep learning based methods, including DeepEdge, HED, RDS [24], RCF) and so on.

The results of the evaluation are shown in the Figure 5. The human eye's performance in the edge detection is called the 0.803-point measurement. The single-scale and multi-scale (MS) versions of RCF have better results than humans. It is also an improvement over the current best algorithm.

- **TABLE 1. The comparison with some competitors on BSDS500 dataset.**

Method	ODS	OIS
Canny [18]	0.611	0.676
gPb-UCM [5]	0.729	0.755
SE [6]	0.743	0.763
DeepContour [10]	0.757	0.776
HED [11]	0.788	0.808
RDS [24]	0.792	0.810
RCF [13]	0.811	0.830
OURS	0.818	0.836

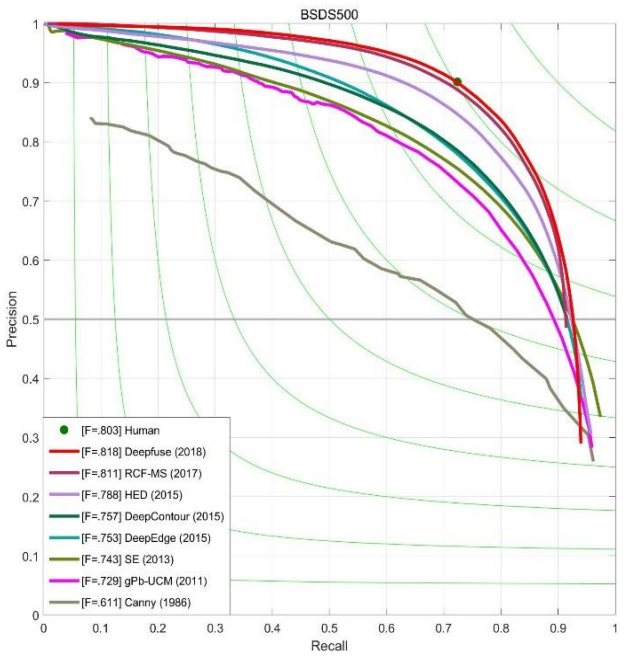


Figure 5. The evaluation results on standard BSDS500 dataset.

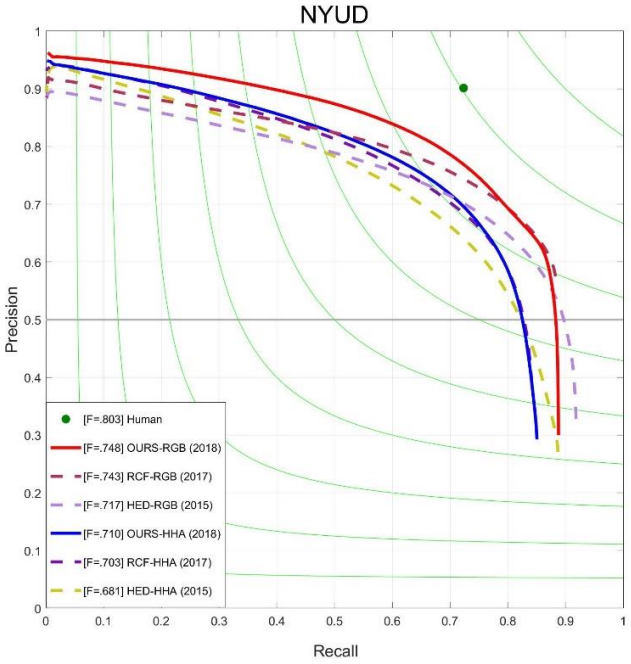


Figure 6. The evaluation results on standard NYUD dataset

4.2 NYUD Dataset

The NYUD dataset consists of 1449 closely labeled aligned RGB and depth images. The NYUD data set was divided into 381 trainings, 414 verifications, and 654 test images. We set up and train our network according to their settings, use the training and validation set, and use full resolution as in HED and RCF.

In both HED and RCF, the re-edited depth map HHA image and RGB image are used respectively in the training process. We also trained HHA images and RGB images separately. The results of training using both images were compared, while other Corresponding methods have higher F-values. From the table 2, we can see that we have improved the ODS F-measure and OIS F-measure by 1.6% and 1.8% on the RGB image, compared to RCF without fusion algorithm, ODS F-measure and OIS F-measure are 3.7% and 4.2% higher on the HHA image. This shows that our method has a competitive performance on this type of problem.

- **TABLE 2. The comparison with some competitors on NYUD dataset.**

Method	ODS	OIS
OEF [23]	0.651	0.667
gPb-UCM [5]	0.631	0.661
SE [6]	0.695	0.708
HED-HHA [11]	0.681	0.695
HED-RGB [11]	0.717	0.732
RCF-HHA [13]	0.705	0.715
RCF-RGB [13]	0.729	0.742
OURS-HHA	0.710	0.722
OURS-RGB	0.748	0.766

5. CONCLUSION

In this paper, a new convolutional neural network is proposed to solve the edge detection problem of natural scene images. We have found that the convolution features gradually become thicker and the middle layer contains many useful details. The novel network structure uses the idea of image fusion to improve the utilization of image feature maps. The proposed network has a better detecting ability on the small edges in the picture, and the accuracy of the detection results on both public data sets is improved, which proves the feasibility of the method. In the future work, we will apply the proposed method for solving related issues such as salient object detection.

6. REFERENCES

[1] Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: CVPR. (2016).

[2] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE TPAMI, 2016.

[3] Shen, W., Zhao, K., Jiang, Y., Wang, Y., Zhang, Z., Bai, X.: Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In: CVPR. (2016) 222–230.

[4] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Generalized boundaries from multiple image interpretations. IEEE TPAMI, 36(7):1312–1324, 2014.

[5] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. IEEE TPAMI, 33(5):898–916, 2011.

[6] P. Dollar and C. L. Zitnick. Fast edge detection using structured forests. IEEE TPAMI, 37(8):1558–1570, 2015.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In IEEE CVPR, pages 1–9, 2015.

[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, pages 91–99, 2015.

[9] G. Bertasius, J. Shi, and L. Torresani. DeepEdge: A multiscale bifurcated deep network for top-down contour detection. In IEEE CVPR, pages 4380–4389, 2015.

[10] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. DeepContour: A deep convolutional feature learned by positivesharing loss for contour detection. In IEEE CVPR, pages 3982–3991, 2015.

[11] S. Xie and Z. Tu. Holistically-nested edge detection. In IJCV. Springer, 2017.

[12] W. Yupei, Z. Xin, and K. Huang, "Deep crisp boundaries," in CVPR, 2017.

[13] Y. Liu, X. H. Ming-Ming Cheng, K. Wang, and X. Bai. Richer convolutional features for edge detection. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3000–3009, 2017.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE CVPR, pages 770–778, 2016.

[16] G. S. Robinson. Color edge detection. Optical Engineering, 16(5):165479–165479, 1977.

[17] I. Sobel. Camera models and machine perception. Technical report, DTIC Document, 1970.

[18] J. Canny. A computational approach to edge detection. IEEE TPAMI, 8(6):679–698, 1986.

[19] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu. Statistical edge detection: Learning and evaluating edge cues. IEEE TPAMI, 25(1):57–74, 2003.

[20] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE TPAMI, 26(5):530–549, 2004.

[21] A. A. Goshtasby. Fusion of multi-exposure images. Image and Vision Computing, 23(6):611–618, 2005.

[22] Prabhakar K R, Srikar V S, Babu R V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs[J]. 2017:4724-4732.

[23] S. Hallman and C. C. Fowlkes. Oriented edge forests for boundary detection. In IEEE CVPR, pages 1732–1740, 2015.

[24] Y. Liu and M. S. Lew. Learning relaxed deep supervision for better edge detection. In IEEE CVPR, pages 231–240, 2016

Optimisation of Feature Space for People Detection from TopView on Light Embedded Platform

Doney Alex
Capillary Technologies Bangalore
doney.alex@capillarytech.com

Prashant Maheshwari
Capillary Technologies Bangalore
prashant.m@capillarytech.com

Sumandeep Banerjee
Capillary Technologies Bangalore
sumandeep.banerjee@capillarytech.com

Subrat Panda
Capillary Technologies Bangalore
subrat.panda@capillarytech.com

ABSTRACT
Detecting people from a top view video feed is comparatively a more difficult problem than pedestrian detection especially on a real time system, as the perceived 2D shape of person from the overhead view is of widely varying nature due to perspective distortions with almost no unique discernible outline shape. Therefore the amount of information available from overhead view is very less. Also, there is a lot of rotational variation of the data from top view. However the advantage of having such a detector is that for a tracking application there would be little occlusion. Hence, this problem is worthy of specialised attention. In recent times there are many deep learning based approaches, but all of them are computationally expensive. We present a method to effectively train a computationally light AdaBoost classifier based detector, which uses the limited amount of information and can give a high accuracy running on a light embedded platform such as Raspberry Pi 3B.

CCS Concepts
•**Computing methodologies → Object detection; Boosting; Feature selection.**

Keywords
Object detection, Boosting, Real-time systems

1. INTRODUCTION
There are many existing algorithms for people detection. Most of them solve the pedestrian detection problem where people are detected from a near horizontal view. Detecting people from a top view was always a difficult problem compared to pedestrian detection as the amount of information available is very limited in top view compared to a horizontal view. Also, there is a lot of rotational variation of the data from top view, along with almost no unique discernible shape. Multiple people in the same frame can lead to challenging problems. Separation of humans close to each other is one example.

*ICV*IP 2018, December 29-31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301541

In this context, robust people detection and separation becomes an extremely difficult task especially in a crowded scene. Illumination changes across scenes also hinder the detection, especially for conventional background subtraction based methods. Therefore this problem requires very specialised attention to solve.The advantage of top view detection especially in cluttered environments is little occlusion and less variations in size with a fixed height of installation of camera.

There are lot of applications for a top view people detector. It may vary from tracking people in a surveillance system or commercial usage like tracking customers in a retail store. But the key is that for a tracking system which needs to be in real time, the detections need to be fast. This is a challenging problem especially if the system is running on an embedded platform with limited computational power. This makes it impossible to use complex deep learning based algorithms which may give high accuracy but need much higher computational power.

We present a method to train an aggregate channel feature based AdaBoost classifier [8, 9] for a top view detection in a way that all the information from top view, though limited is effectively used and can run on a light embedded platform at a high speed. We introduced multiple hyper parameters and an iterative refinement hard negative mining training method for training the classifier that when it is integrated as a detector, achieved a high accuracy even when there were multiple people side by side. The major contributions of our method are: (1) In contrast to many other approaches (which are not deep learning based), our method does not use foreground/background segmentation. Therefore it does not suffer from problems related to background modelling like a static person becoming part of background, illumination changes etc. (2) We introduce two hyper parameters for training, C called crop-factor and S called resize factor. C improves the detections when there are people close to each other in a scene and S helps in optimising the detector for speed. (3) We introduce an iterative refinement hard negative mining training method which improves the accuracy of the AdaBoost classifier.

For experiments and evaluation, the classifier was integrated to a multi-scale detection system [9] with a tracker(Kalman algorithm [16] with Hungarian Assignment [19]), capturing video feed at 214 x 160 resolution and deployed in a retail store environment on an embedded platform, so that the number of people coming into the store and going out of the store can be counted in real time. We performed the same experiments on recorded videos also. The system gave 96% accuracy in real time owing to the fast and accurate detector.

2. RELATED WORK

Most of the people detection algorithms solve the pedestrian detection problem. A detailed study has been done by P Dollar *etal*[11]. The advantage of the horizontal view is that it is more stable view with more detailed view of the features with negligible rotational variance. But for an application which requires tracking, this may fail because of possible occlusion and variations in size of the person with the variation of distance from the camera. One of the notable work in frontal (horizontal view) detection is HOG-SVM based classifier [5]. The HOG-SVM and its modifications [30][23]work well on pedestrian detection but for a top view detection the amount of information available from overhead view is very less as there is only a circular blob of head which is visible and hence the gradient is very limited. Along with the head, partial face is also visible at some angles but it is not sufficient to extract and work on facial features. Besides the face is visible for a very short duration in the frame for a moving person. Another approach is using multiple channel features. The integral channel features presented by P Dollar *et al*[10] has a higher dimensional feature space which can extract much more information compared to HOG [5].

As mentioned, top view detection is a much complicated problem. Many methods have been proposed to tackle the problem of top view detection. Using a simple HOG based detector [2] similar to that in pedestrian detection [5] will not give accurate results because of the lack of availability of information from top view. Pure vision based approach [4, 27] like background (BG) subtract, blob detection etc are susceptible to lighting conditions, shadows, reflections, movement of random objects etc. Methods which involve background modelling are sensitive to illumination changes, especially for colour data captured by traditional colour cameras, since the pixel values can change drastically in consecutive frames over a short period of time due to the sensitivity of camera to lighting conditions. Blob detection using BG subtract fails whenever two or more people are in close proximity. Multiple persons may get counted as one. The foreground segmentation has lot of shortcomings. For *e.g* a person that stands still for a while eventually gets integrated into the background. Similarly, the shadows in the scene also get detected as foreground blobs, although this can be improved by shadow detection and elimination. There exist methods using multiple cameras [25-27, 29] or using complex features [22]. The cost of multiple cameras and processing power required makes the use of these methods undesirable especially on a light embedded platform. Cost and complexity are again the factors in using time of flight camera [18].

Recently deep learning methods have shown to achieve significant improvement in detections compared to all conventional methods. There exist many CNN networks like YOLO [24], Mobilenet based SSD [13, 17] and lot more [20, 21] or dedicated hardware which can run deep learning networks like Intel-Movidius NCS [14], which can do detections with very high accuracy. These networks can be retrained to solve top view people detection and can achieve high accuracy. Even though they can give better accuracy, the computational requirement is high and cannot run on real time on a light embedded platform as shown in various benchmarks studies [6, 12, 15]. We want a system which can run on a light embedded platform like Raspberry Pi 3B and can achieve a very high speed (close to 40 fps). Hence deep learning based detectors are not the right choice for such a system.

AdaBoost classifier is a common choice of classifier for real time systems. Q Zhu et *al*[30] showed how to feed HOG features to a cascaded AdaBoost classifier for pedestrian detection. Viola and Jones [28] showed how to train a cascaded AdaBoost classifier using integral image features for object detection. There are other notable works also [1][3], which use cascaded AdaBoost classifier for human detection. In order to use the AdaBoost classifier trained as detector, the classifier developed should search the entire image on multiple scales for detection. There are multiple methods for optimising the multi-scale search [8]. In an over head detection system with a fixed camera height the number of scales required can be limited as the size does not vary much. In our work, we used aggregate channel features for overhead detection. The novelty is how to optimally use these features and train AdaBoost classifier to achieve maximum accuracy in top view people detection and tracking in real time.

3. TRAINING DATA GENERATION AND MODEL TRAINING

We implemented a modified C++ version of AdaBoost classifier in [7]. The classifier was trained on 3 million positive and negative samples.

3.1 Generation and Augmentation of Positive Training Samples

Augmentation of data samples is a common practice during training in most machine learning algorithms. It is significant, especially in cases where number of training samples available is limited. In those cases there will be some available samples and we augment them to increase the total number of training samples. In our proposed methodology, we went one step further to create the entire positive data set artificially from few raw images. From a large set of images containing the overhead views of people, we begin with carefully curating the samples with binary annotation, with foreground pixels opaque and background pixels transparent Fig 1(a). The sample images may contain one or more people, which are duly annotated by rectangles marking their location. We generate the positive training samples in two steps. First, to provide rotational invariance to our detection algorithm, we generate a set of rotated samples by extracting the annotated rectangle and giving it planar rotation at a number of angles (usually 8 to 12 variations) covering the whole 360 degrees. Secondly, to maintain statistical robustness, we combine the rotated samples with a set of random backgrounds Fig 1(b). To further improve generality, we use a large set of backgrounds but randomly choose a much smaller subset of them for each sample image. We combine the opaque foreground with pixels of the background from the transparent pixel regions of the samples. We use pyramid blending to minimise the effect of boundary artifacts at the edge of foreground and background like in Fig 1. This process allows us to synthetically generate an incredibly large number of training samples from a relatively smaller set. Also, it helps to cover a large range of environment and orientation variations, which otherwise would be extremely cumbersome to create and directly capture. We combined about 13K foreground images combined with 1K background images to generate 1M positive data samples. This large number of images covers variation of clothing, hair colour, etc.

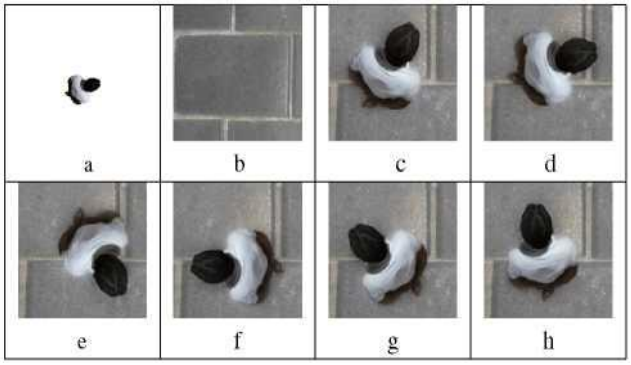


Figure 1. Positive data generation and augmentation : (a) Foreground opaque and background transparent image. (b) A random background. (c - h) Foreground blended on background in different angles.

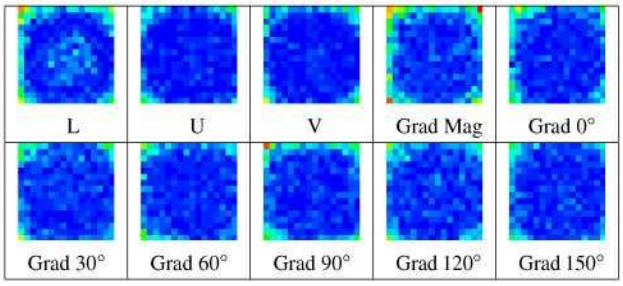


Figure 2. Importance map of feature vectors from training example for all 10 channels(L,U,V, Gradient Magnitude and Gradient Histogram in 6 different angles). Red (most important) to Blue(least important).

(*Wd* = *Ht* = 128 , C = 16 and k = 1)

3.2 Negative Samples

We take a large set of images of environments where the system is expected to be deployed and select the images which do not contain any part of people to be detected. From this set of selected images, we perform random sample generation. For this we create image pyramids out of each selected negative image, and at each scale we randomly pick a number of sub-images of same resolution(as positive samples) with a given probability. This way we generate a large set of negative samples covering a varied range of negative backgrounds which are also robust to scale/size variations.

3.3 Feature Space

We selected aggregate channel features to describe the object. Thefirst step is to extract the channel features from each generatedsample. We have selected the channels such as L, U, V, GradientMagnitude, Histogram of Gradients along 0◦, 30◦, 60◦, 90◦, 120◦,150◦. For each sample at resolutionW x H, we compute N channelswith a sub-sample factor k. Number of features generated is Fn givenby

$$F_n = \frac{WHN}{k^2} \quad (1)$$

We flatten the features into a vector of the dimension F_n. This id done for all positive and negative samples.

3.4 AdaBoost Classifier Training

We have used a multi-staged training of AdaBoost classifier to perform the task of object classification. The algorithm is based on training a set of weak classifiers applied in sequence order to incrementally weed out the negative samples based on different decision boundaries, while preserving the positive samples. The first classifiers reject very few negative samples, whereas passing almost all positive samples. The subsequent weak classifiers use different criteria to remove the negatives that have passed through by earlier weak classifiers, while maintaining the clearance rate of positive samples. Each classifier is a depth limited decision tree, with nodesbased on different features picked for maximal separation boundary of positive and negative samples. All features might not be equally important. Fig 2 is a histogram plot of the entire feature space describing how many times the particular feature is used by the trained classifier model. The classifier was trained on a data-set created as mentioned in 3.1 and 3.2, which contained 1 million positive samples and 2 million negative samples.

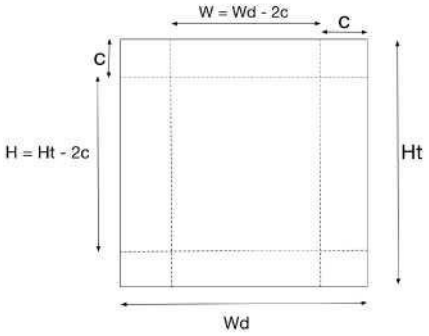


Figure 3: Crop Factor explained.

4. OPTIMISATION OF FEATURE SPACE

Object Detection entails determining location and size of target object in a given image. Correct localisation and scale is essential to detection performance and especially important if the system incorporates tracking of object over image sequences. We have observed that location, size, accuracy and precision of the detection can be influenced by altering the way we interpret the feature vectors computed for model training. We propose a data driven approach to determine the optimal feature space from the given samples by selecting a subset of image samples for feature extraction. The samples are generated by us at a higher resolution and cropped/resized for fine tuning for best detection quality.

4.1 Crop Factor Optimisation

From the overhead view, only the top of the head and shoulders are visible for people passing beneath the camera. For our system to learn how to detect people from overhead view, we first trained it

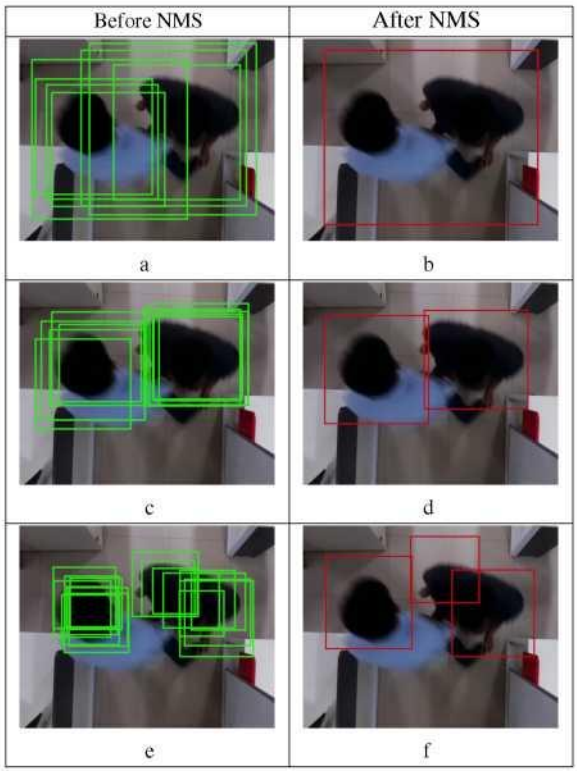


Figure 4. Effect of crop-factor C **: For a & b,** C **= 0; For c & d,** C **= 12; For e & f,** C **= 24; (** Wd= *Ht* = 128 **)**

with annotating the complete person as visible from the top. We were able to get a good convergence on the model and detect people well. But the problem arose with multiple people in close proximity. This is due to merging of detected rectangles during the Non-Maxima Suppression (NMS) step. In multi-scale object detection, we can have detections at more than one scale. To avoid detecting the same object multiple times, detections overlapping more than a certain predefined ratio are combined to represent one consolidated detection rectangle. But for two persons appearing in the scene side by side may give detections that are overlapping with each other as shown in Fig 4(a). In this scenario, NMS merges the two rectangles into one thereby leading to incorrect detection.

We solve the above mentioned problem due to NMS by introducing a new hyperparameter C called crop-factor. For data generated as mentioned in section 3.1 (with size Wd x Ht), we select only the centre part of the sample image (with size W x H) as shown in Fig 3. Depending on the crop factor, this removes part or whole of the shoulder from the sample, keeping only the top view of the head. By doing so, the detector starts detecting and marking only the head portion thereby reducing or eliminating the overlap between detections of two persons in close proximity consequently increasing the detection quality. But, this does not come without its own caveats. By removing the shoulders wholly, we reduce the training sample (the head) to little more than roughly a round object. This causes loss of useful distinct shape features causing loss of trained model generality.

Excessive cropping may lead to an increase in the rate of false positive detections. Less or no cropping leads to bad object localisation whereas over cropping leads to loss of generality of model as shown in Fig 4. The system (classifier training + detection) needs to be fine tuned using a data driven approach to strike the right balance. We need to find the optimal cropping factor to determine the best way to avoid side effects while maintaining sufficient model richness.

4.2 Sample Size Optimisation

After cropping the samples, we use another hyper parameter S called resize factor to resize the sample to a suitable size for feature space extraction. The effective size of the sample will be W/S x H/S. The resize factor S is different from k in equation 1.*k* represents sub sampling in feature space whereas S is the parameter for resizing the image sample before computing the features. In all our experiments k was set as 2. When the classifier is used as a detector by sliding and searching through the input image at different scales as in [9], the input image can be resized to reduce the number of features to be calculated. But when we run the detector, effective step/stride while sliding will be scaled up by S which will lead to lose of some detections in some scales. Hence we resize the samples while training. For detection, the sliding is done in default steps(configured by the detector) but the samples fed for feature generation are subsampled. For large values of S, we get a very fast detector owing to less number of features to compute, but, it leads to poor detection due to increase in false detections. By reducing S, we tend to get increasingly better accuracy. But, due to processing constraints on the given computing platform, the computation time keeps increasing. There is a sweet spot for resized sample size that gives us the desired detection accuracy at optimal frame rate. This is determined using data driven fine tuning of the system.

4.3 Iterative Refinement Hard negative mining

We perform a multistage training of the AdaBoost classifier. We begin with a stage with low number of weak classifiers. We have a collection of top view images expected to be seen by the overhead camera without any persons present. After training classifiers of one stage, we use it to look for false detections in this collection of negative background images. If any detections are found(which are actually false), the sub-images corresponding to the rectangular blocks are added to the negative sample set for the training model to learn. After each round of negative sample mining, we increase the number of classifiers and train the next stage. We increased the number of weak classifiers by a factor of 4 after each stage. We proceeded with a four stage system with 32, 128, 512,2048 classifiers. With each new stage, out false positive rates drop. The stopping criteria is either convergence of true positives or cap on number of classifiers in one stage.

The classifier determines the samples as positive with a confidence value, which can vary from negative to positive. True negatives generally give negative confidence value, false positives are low positive values, and true positives give a relatively higher positive value. Determining the confidence threshold to clearly separate the true positives from the rest is a matter of fine tuning and trade off. By proper use of the hard negative mining, we are able to push the effective confidence threshold closer to zero irrespective of usage situation. This makes our detector more generic and robust. Iterative Refinement Hard negative mining also acts as a control knob for over fitting and regularisation. We can overfit by adding a lot of top view images of scenes specific to location of deployment. We can avoid extreme over fitting by having a lot of generic top view scenes with out people in it. In our case, we wanted to deploy the system in retail stores for counting the number of people coming in and going out of the stores. We added a few images of the background scenes captured from top view from different stores to the collection, which improved the accuracy.

We tune the hyper parameters by iterative training mechanism. We

evaluate different accuracies by various experiments explained in the section 5 and feed back is given for next iteration for optimisation. The entire training mechanism is explained in Fig 5.

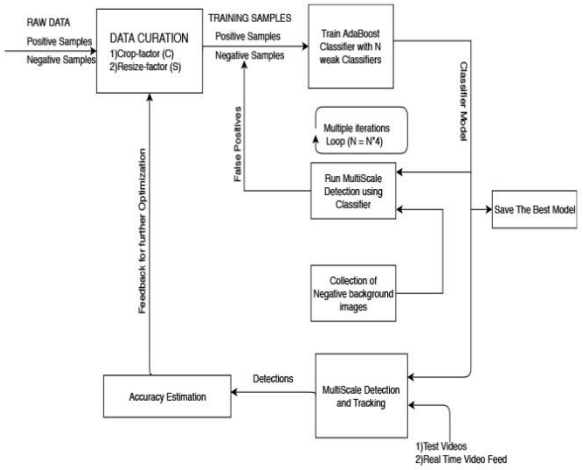


Figure 5. Complete Training Process.

5. EXPERIMENTS AND RESULTS

The entire system was coded in C++ and was deployed in a Raspberry Pi 3B. Leaf level functions were coded in ARM SIMD (NEON) for optimisation. For experiments and evaluations we recorded videos of people walking in multiple directions from top view, from various heights (8f t to 11f t). We tried to cover various scenarios of multiple people walking together, multiple people crossing etc. The videos were recorded at constant frame rate (20 fps). We conducted experiments to evaluate three metrics. 1) Classification Accuracy (Ca): This is the accuracy of the trained AdaBoost binary classifier. We used artificially generated samples during training but for evaluating Ca, we use annotated samples from true images obtained from the recorded videos. We define Ca as and Negatives. 2)Detection Accuracy(D_a): We integrate the trained classifier as detector by scanning window at different scales as in [9]. As the variation in height is limited we restricted the number of scales to 4 and step in sliding is kept as 1. The Detection Accuracy is evaluated using annotated images and we measure IOU (with .6 as threshold). 3)Counting Accuracy(P_a): The detector was integrated to a tracker (Kalman filter [16] with Hungarian assignment [19]). The objective was to count the number of people going in and out. One direction was defined as 'In' and its opposite was defined as 'Out' and adjacent sides were split equally. The direction of the person was decided based on which direction he came into the frame and which direction he went out. We evaluated two Counting Accuracies, using the recorded videos (P_{av}) and another using real time video feed(P_{ar}). The difference between P_{av} and P_{ar} is that P_{av} runs on videos recorded at constant fps of 20, whereas in real time deployment a buffer mechanism is provided to form a temporary queue such that the system runs on a frame rate that it can achieve dynamically based on the speed of the detector. The optimisations done on the classifier to improve the evaluation speed comes into picture here and higher frame rate will improve tracking resulting in a better Counting Accuracy. We define a metric Counting AccuracyP_a

$$C_a = \frac{T_p + T_N}{P + N} \tag{2}$$

where Tp and TN represent number of True Positives and True Negatives respectively, P and N is the total number of Positives

where$In_{true}$and $Out_{true}$is the actual number of people going In and Out, $In_{sys}$and $Out_{sys}$is the count given by the system.

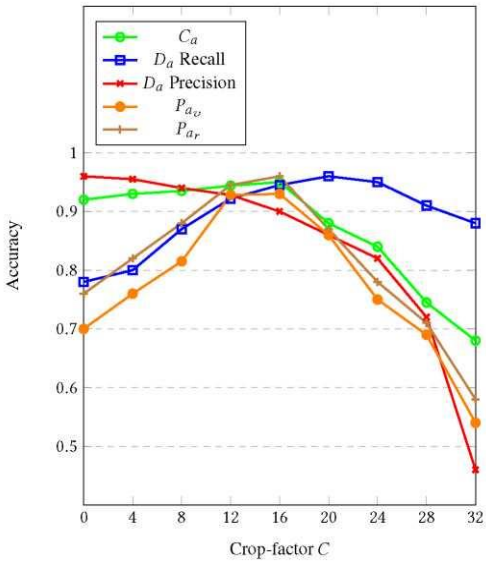


Figure 6: Variation of Classification Accuracy, Detection Accuracy and Counting Accuracies with Crop Factor. *Wd = Ht* **=128, S = 3/2**

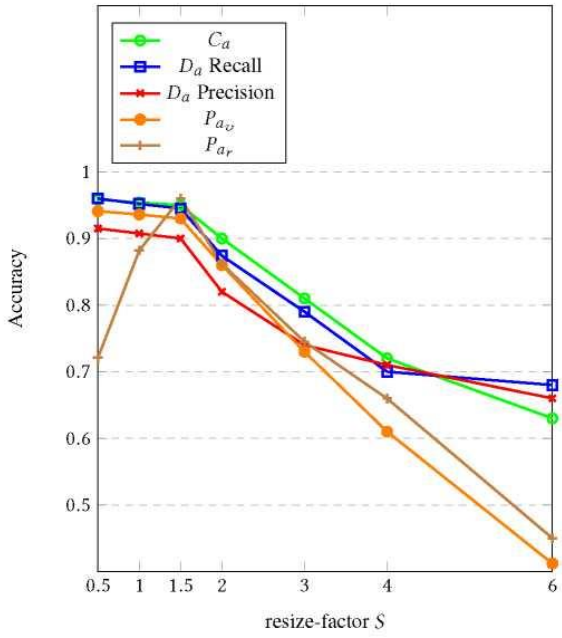


Figure 7: Variation of Classification Accuracy, Detection Accuracy and Counting Accuracies with resize-factor.

(Wd = *Ht* **= 128, C = 16)**

$$P_a = \frac{In_{sys} + Out_{sys}}{In_{true} + Out_{true}} \tag{3}$$

The presented algorithm is optimised for overhead view, to perform well on low compute-power devices. Most datasets available for overhead view are not perfectly top view. Most of them are at an angled view where partial face / side of head can be seen. Our system is designed to handle an exact overhead view. Since datasets of this type of overhead view are not publicly available for benchmarking, comparative benchmarking cannot be done in a straightforward manner. The experiments for evaluating P_{av} is done on prerecorded videos. We used 1279 videos which were 20 minutes each containing a total of 11946 bi-directional count. These videos were recorded at 20fps and ground truth is calculated by visually counting people movement in videos. This is compared with system output to generate accuracy numbers. The P_{ar} was evaluated by installing devices at 110 different locations. Here the system captures frame at dynamic rate which is based on the processing time taken by each frame. Higher frame rate (more than 20 FPS) gives better tracking accuracy and hence better counting accuracy. Ground truth count is maintained by a person, monitoring the IN/OUT count of the camera monitored area. At the same time our device is also counting. Later the count from two sources is compared (assuming person's count as ground truth) to compute accuracy. This is the real time evaluation of algorithm. In both and P_{av} and P_{ar}, the algorithm was exposed to different illumination and lighting conditions such as - indoor, outdoor, direct sunlight, indirect sunlight, day, evening, and night time and different installation heights. Other experiments for C_a and D_a were conducted using annotated images obtained from part of prerecorded videos mentioned above/

The variations of accuracies with crop-factor C is shown in Fig 6. We observe that initially the Classification Accuracy C_a does not change much with increase in crop-factor but with further increase in C, the Classification Accuracy falls. This is due to loss of distinct shape features causing loss of trained model generality. Further we see that Detection Accuracy recall increase with increase in crop factor. This is because of the improved performance of the detector in cases where multiple people appear side by side and there is limited overlap of detection rectangles and hence not lost during NMS. However the detection precision drops with increase in the crop factor due to the increase in false positives in lower scales. Because of these, we observe that the Counting Accuracies P_{av} and P_{ar} initially increase with crop factor, but excess cropping may lead to a decrease in Counting Accuracy. So in order to use the detector for an application that involves tracking, we need to select an optimal crop factor for best performance. Both P_{ar} and P_{av} show similar patterns but P_{ar} is slightly better because while evaluating P_{ar}, the system is able run on a higher frame rate(close to 40 fps), which resulted in a better tracking accuracy which further improved the counting.

The impact of resize-factor S on accuracies is shown in Fig 7. We observe that C_a and D_a tend to decrease because with increase in S, there is lack of sufficient features for correct classification. S acts as an optimisation for processing time, which results in an improved frame rate. For evaluation of improvement in computation requirement with S, we ran the classifier on a Raspberry Pi 3B platform in single threaded form and computed the time taken by the classifier. The time taken by a positive classification will be always higher as it has to pass through all the weak classifiers. Hence, we chose to measure the time taken for a positive classification and the results are shown in Table 1.

Table 1. Computation time for classifier with resize-factor S. (Wd = Ht = 128, C = 16)

Scale Factor S	Time taken for 100 samples (in milli sec)
1/2	11.286
1	3.44
3/2	1.7
2	0.922
3	0.478
4	0.321
6	0.29

Even though the C_a and D_a decrease with S, we observe that the P_{ar} increases initially but drops later.This is because, with the decrease in computation time required for a detection, the system is able to work on a higher frame rate which improved tracking, resulting in better counting. But a further increase in S made the classification accuracy

Table 2. Effect of Hard negative mining. (Wd = Ht = 128,C =16, S = 3/2)

	Without Hard negative mining	With Hard negative mining
C_a	0.89	0.95
D_aRecall	0.9	0.945
D_aPrecision	0.87	0.9
P_{av}	0.87	0.930
P_{ar}	0.91	0.96

to drop, which further decreased the $P_{ar.}$ The resize-factor did not improve P_{av} because the frame rate is fixed. S has an impact when the system is deployed in real time and where a decreased computational time can contribute in improving the accuracy. So choosing the optimum resize factor is the key to obtain higher accuracy.

The impact of Iterative Hard negative mining is shown in Table 2. We added a few negative background images of the scene in which the system was deployed along with a lot of background images. The improvement was reflected as a reduction in number of false positives which improved all accuracies.

6. CONCLUSION

Detecting people from top view is a complex problem but has a lot of application, especially if we add a tracker to it. We proposed a method to train an AdaBoost classifier which can be deployed on a light embedded platform with limited computational capability. We introduced two hyper parameters, C which improved detection accuracy where there were multiple people in close proximity and S which reduced the computation time required for the detector. We also introduced an iterative refinement hard negative mining training method which improved the classification accuracy. Though modern deep learning based methods may achieve better accuracy, the minimal computational requirement and simplicity is what makes the proposed method desirable.

Future work will include a more exhaustive experimental work to improve the detection accuracy in a more robust outdoor environment, so that that the system can be used for a surveillance sort of application.

REFERENCES

[1] J. Begard, N. Allezard, and P. Sayd. 2008. Real-time human detection in urban scenes: Local descriptors and classifiers selection with AdaBoost-like algorithms. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 1-8. https://doi.org/10.1109/CVPRW.2008.4563061

[2] Ben Benfold and Ian Reid. 2009. Guiding Visual Surveillance by Tracking Human Attention. In *Proc. BMVC*. 14.1-14.11. doi:10.5244/C.23.14.

[3] P. Cerri, L. Gatti, L. Mazzei, F. Pigoni, and H. G. Jung. 2010. Day and night pedestrian detection using cascade AdaBoost system. In *13th International IEEE Conference on Intelligent Transportation Systems*. 1843-1848. https://doi.org/10.1109/ITSC.2010.5625019

[4] I. Cohen, A. Garg, and T. S. Huang. 2000. Vision-based overhead view person recognition. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Vol. 1. 1119-1124 vol.1. https://doi.org/10.1109/ICPR.2000.905668

[5] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 886-893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[6] Xiaofan Xu Dexmont Pena, Andrew Forembski and David Moloney. 2017. Benchmarking of CNNs for Low-Cost, Low-Power Robotics Applications. (2017). http://juxi.net/workshop/deep-learning-rss-2017/papers/Pena.pdf

[7] P Dollar. [n. d.]. Piotr's Computer Vision Matlab Toolbox (PMT). ([n. d.]).https://github.com/pdollar/toolbox

[8] P. Dollar, R. Appel, S. Belongie, and P. Perona. 2014. Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 8 (Aug 2014), 1532-1545. https://doi.org/10.1109/TPAMI.2014.2300479

[9] Piotr Dollar, Serge Belongie, and Pietro Perona. 2010. The Fastest Pedestrian Detector in the West. In *Proc. BMVC*. 68.1-11. doi:10.5244/C.24.68.

[10] Piotr Dollar, Zhuowen Tu, Pietro Perona, and Serge Belongie. 2009. Integral Channel Features. In *Proc. BMVC*. 91.1-91.11. doi:10.5244/C.23.91.

[11] P. Dollar, C. Wojek, B. Schiele, and P. Perona. 2009. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 304-311. https://doi.org/10.1109/CVPR.2009.5206631

[12] DT42. 2017. Run Object Detection using Deep Learning on Raspberry Pi 3(1). (2017). https://medium.com/dt42/run-object-detection-using-deep-learning-on-raspberry-pi-3-1-55027eac26c3

[13] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (04 2017).

[14] Intel. 2017. Intel Movidius Neural Compute Stick. (2017). https://developer.movidius.com/

[15] Sarthak Jain. 2018. How to easily Detect Objects with Deep Learning on Raspberry Pi. (2018). https://medium.com/nanonets/how-to- easily- detect- objects-with- deep- learning- on-raspberrypi- 225f29635c74

[16] Rudolph Emil Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME-Journal of Basic Engineering* 82, Series D (1960), 35^5.

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision - ECCV 2016*, Bastian Leibe, Jiri Matas, NicuSebe, and Max Welling (Eds.). Springer International Publishing, Cham, 21-37.

[18] Carlos A. Luna, Cristina Losada-Gutierrez, David Fuentes-Jimenez, Alvaro Fernandez-Rincon, Manuel Mazo, and Javier Macias-Guarasa. 2017. Robust people detection using depth information from an overhead Time-of-Flight camera. *Expert Systems withApplications* 71 (2017),240-256. https://doi.org/10.1016/j.eswa.2016.11.019

[19] James Munkres. 1957. ALGORITHMS FOR THE ASSIGNMENT AND TRANS- PORTATIONPROBLEMS. (1957).

[20] Vinod Nair, Pierre-Olivier Laprise, and James J. Clark. 2005. An FPGA-based People Detection System. *EURASIP J. Appl Signal Process.*2005(Jan.2005),1047-1061. https://doi.org/10.1155/ASP.2005.1047

[21] AznulQalid Md Sabri NouarAlDahoul and Ali Mohammed Mansoor. 2018. RealTime Human Detection for Aerial Captured Video Sequences via Deep Models. (2018). https://doi.org/10.1155/2018/1639561

[22] O. Ozturk, Toshihiko Yamasaki, and KiyoharuAizawa. 2009. Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. 1020-1027. https://doi.org/10.1109/ICCVW.2009.5457590

[23] YanweiPang,YuanYuan,XuelongLi, andJingPan.2011. EfficientHOGhuman detection. *Signal Processing* 91, 4 (2011), 773 - 781. https://doi.org/10.1016/].sigpro.2010.08.010

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision andPatternRecognition (CVPR)*. 779-788. https://doi.org/10.1109/CVPR.2016.91

[25] Thiago T. Santos and Carlos H. Morimoto. 2011. Multiple camera people detection and tracking using support integration. *Pattern Recognition Letters* 32,1 (2011), 47 - 55. https://doi.org/10.1016/j.patrec.2010.05.016 Image Processing, Computer Vision and Pattern Recognition in Latin America.

[26] T. E. Tseng, A. S. Liu, P. H. Hsiao, C. M. Huang, and L. C. Fu. 2014. Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras. In *2014IEEE/RSJInternational Conference on IntelligentRobots and Systems*. 4077-4082. https://doi.org/10.1109/IROS.2014.6943136

[27] TimvanOosterhout, SanderBakkes, andBenKrAfise. 2011. HEAD DETECTION IN STEREO DATA FOR PEOPLE COUNTING AND SEGMENTATION. InProceedings of the International Conference on Computer Vision Theory and Applications - Volume 1: VISAPP, (VISIGRAPP 2011). *INSTICC, SciTePress, 620-625.* https://doi.org/10.5220/0003362806200625

[28] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001,* Vol. 1.1—511—I—518 vol.1. https://doi.org/10.1109/CVPR.2001.990517

[29] ZhognchuanZhangandF. Cohen. 2013. 3D pedestrian tracking based on overhead cameras. In *2013 Seventh International Conference on Distributed Smart Cameras (ICDSC).* 1—6. https://doi.org/10.1109/ICDSC.2013.6778235

[30] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. 2006. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR '06).* IEEE Computer Society, Washington, DC, USA, 1491—1498. https://doi.org/10.1109/CVPR.2006.119

Chapter 3: Image 3D Reconstruction

3D Face Reconstruction from Low-Resolution Images with Convolutional Neural Networks

Rouven Winkler[1], Chengchao Qu, Sascha Voth, Jürgen Beyerer[2]
Fraunhofer Institute of Optronics, System Technologies, and Image Exploitation (Fraunhofer IOSB)
Fraunhoferstr. 1, 76131 Karlsruhe, Germany
chengchao.qu@iosb.fraunhofer.de

ABSTRACT

During the past years, convolutional neural networks (CNNs) have widely spread as a powerful tool for tackling a variety of challenges posed in computer vision. Consequently, the trend neither does stop at 3D face reconstruction: Recently, several CNN-based approaches for reconstructing the dense 3D geometry of a face from only a single image have been introduced. However, while all of these methods deal with 3D face reconstruction in the high-resolution (HR) case, reconstruction in low-resolution (LR) surveillance scenarios by means of CNNs has not received any attention so far.

With this work, we address that gap, being the first to propose a CNN architecture specifically tailored to LR 3D face reconstruction: We introduce an end-to-end trainable CNN capable of simultaneously estimating 3D geometry and pose of a face given a single LR image. By coupling our network with a state-of-the-art LR face detector, we build a 3D face reconstruction pipeline ready for integration into real-world applications.

We conduct systematic evaluation on LR versions of the in-the-wild AFLW2000–3D dataset, considering decreasing interocular distances (IODs) down to three pixels. The results show superior performance of the proposed method in the LR domain over state-of-the-art approaches, for both 3D face reconstruction and the closely related face alignment task.

CCS Concepts

• Computing methodologies → Biometrics. • Computing methodologies → Neural networks.

Keywords

3D face reconstruction; low-resolution; CNN

1. INTRODUCTION

In times of increasing video surveillance of public places mainly motivated by security considerations, massive amounts of facial image data are collected, with the captured images often exhibiting poor quality due to cheap hardware and wide-angle setups. As a result, automated solutions for processing and analyzing low-resolution (LR) face images are gaining more and more in importance, making especially LR face recognition and face super-resolution (SR) of particular interest. However, various challenges inherent to the LR face recognition task—amongst them, importantly, automated alignment as well as resistance to pose variations—were to be considered as unresolved in the review [23] conducted a few years ago, and this assessment, in substantial parts, appears to be still valid today. By no means, however, does this indicate the absence of attempts to cope with the respective difficulties—in fact, there are several promising approaches, some of which make use of additional geometric information about the face in question: while 3D face shapes are beneficial for solving a variety of computer vision tasks that deal with high-resolution (HR) face image data, the same is true for facial image processing in the LR domain, and consequently, a number of approaches have successfully exploited 3D information for SR [18] and recognition [6] of LR faces. Nevertheless, with existing 3D face reconstruction approaches being designed primarily for processing HR face images, there is barely any method available which is suited for directly reconstructing 3D face shape from LR image data, with the exceptions in [9, 17] for fitting a 3D face model on LR images. However, as admitted in [8], these analysis-by-synthesis frameworks struggle with in-the-wild scenarios.

Convolutional neural networks (CNNs) have proven to be well-suited for solving a variety of different computer vision tasks—it is no exception here. Thus, most state-of-the-art methods are CNN-based. Motivated by the recent success of cascaded regression for face alignment, the first algorithms such as [15, 25] also employ a cascaded architecture to gradually improve the fitting quality. However, they are cumbersome to train, as an end-to-end fashion is not possible. On that account, later approaches tend to utilize a single CNN. Richardson et al. [19] propose a two-stage iterative network for enhancing fine structural details. They leverage synthetic 3D Morphable Model (3DMM) renderings to generate ground truth data for training [20]. Dou et al. [4] also use this strategy for their end-to-end CNN, which is restricted to 3DMM shape only. Direct regression of the dense 3D shape without the intermediate 3DMM abstraction is another popular direction. As an example, a volumetric representation of the face shape is regressed by Jackson et al. [12]. Alternatively, the Face Alignment Network (FAN) [2] addresses 2D and 3D face alignment using heatmap regression. Despite the successes achieved so far for large-pose in-the-wild reconstruction, it is worth noting that no CNN-based solutions for true 3D LR face reconstruction exist to date. The lately introduced Super-FAN [3] basically super-resolves input images prior to alignment.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong.

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301519

[1] This work was done when Rouven Winkler was an intern at Fraunhofer IOSB.

[2] Jürgen Beyerer is also with the Vision and Fusion Laboratory (IES) at the Karlsruhe Institute of Technology (KIT).

In order to address that research gap, we introduce a novel method for LR pose-aware 3D face reconstruction in the wild, which simultaneously estimates the underlying pose and 3D shape of a face presented in a single LR image by employing a new CNN architecture as its core component. Importantly, our approach is fully self-contained: Concretely, by integration of a state-of-the-art face detector for LR faces [10] into our reconstruction framework, there is no dependence on external bounding box prior knowledge for region of interests (ROIs) determination, eliminating the need for providing any information beyond the raw 2D face image itself to our framework and thus making it robust to initialization [24] and suitable for seamless application in real-world scenarios. To the best of our knowledge, this is the very first CNN-based attempt on 3D face reconstruction specifically focusing on LR image data.

Our network builds upon the end-to-end, single-pass architecture proposed in [4], which operates on HR images of 180×180 pixels. However, with LR images carrying limited information, the number of clues exploitable for reconstruction is considerably reduced. Moreover, the depth of the network is limited by the low input resolution as well. Therefore, motivated by the LR recognition network [7], we investigate several back-end candidates for LR feature extraction, and propose a specially designed architecture for LR face reconstruction. Furthermore, an additional branch for pose estimation is introduced, making our network, coined Low-Resolution 3D Face Fitting Network (LR-3DFF-Net), ready to aid in the context of SR or pose-aware recognition tasks.

Our main contributions can be summarized as follows:

- Building on a state-of-the-art CNN architecture for HR 3D face shape reconstruction [4], we develop a novel network specifically tailored to LR scenarios, which is capable of simultaneously estimating both facial shape and pose from a single LR input image.
- We systematically investigate three different back-end types for facial feature extraction by conducting thorough experiments to reveal their impact on the overall reconstruction performance, finally determining the one suited best for our LR-3DFF-Net.
- By coupling with a state-of-the-art LR face detector [10], we form a fully self-contained pipeline for LR 3D face reconstruction on in-the-wild image data, ready for integration into real-world applications.

2. LR-3DFF-NET

We start this section with an overview of our 3D face reconstruction method, outlining the employed pipeline and briefly describing the involved components. After that, detailed information about the network architecture and the training procedure is given.

2.1 Reconstruction Pipeline

LR facial analysis is a complex task, usually consisting of an amalgam of various challenging sub-problems. It is particularly the case for the setup of LR 3D reconstruction. To alleviate the difficulties, we devise a dedicated and pragmatic workflow to ensure robustness on LR images, which is illustrated in Figure 1.

2.1.1 Face Detection

Given an image of low quality from a surveillance camera, the first step is to locate the target face within the image. Existing 3D face reconstruction frameworks either assume provided bounding boxes from preceding stages, or adopt off-the-shelf algorithms, which oftentimes fail with small faces in the image. Hence, it is nontrivial and necessary to integrate a robust detector for our LR pipeline. The state-of-the-art approach by Hu and Ramanan [10], especially appropriate for "finding tiny faces", is without doubt the best option. Thanks to the larger context region and pyramid scale-space, it can precisely detect faces with a size of merely a few pixels, providing stable initializations for our network. This is crucial since regression-based alignment or reconstruction methods are sensitive to small disturbances in the initial bounding boxes [24].

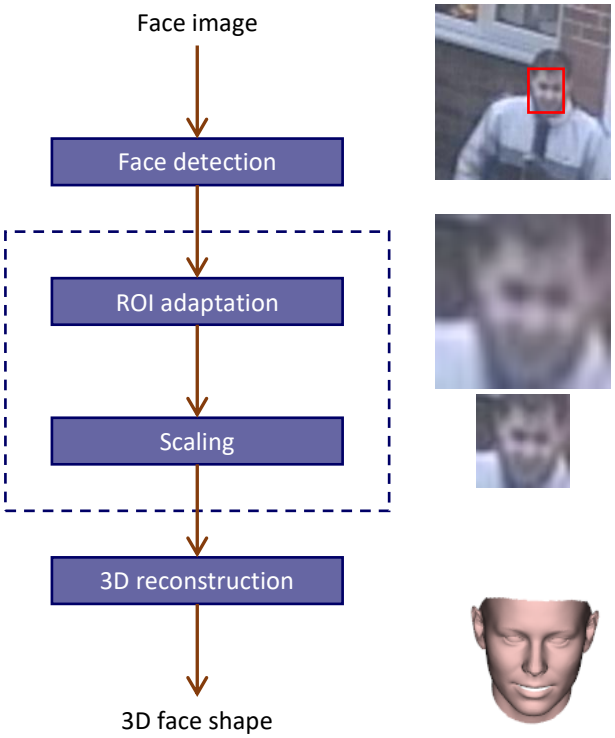


Figure 1. Overview of the reconstruction pipeline.

2.1.2 ROI Adaptation

The tight bounding boxes found by [10] (see the first image in Figure 1) need to go through several adjustments, before they are applicable as the input to our network. The reason is twofold. Firstly, like almost all available approaches in the literature [4, 12, 15, 25], we use quadratic input size. And secondly, the tight margin is detrimental, as extra context knowledge around the face is not able to be included to unleash the power of deep nets.

For a rectangular bounding box B with height H_B and width W_B, its quadratic form Q with side length L_Q, centered at the same position as B, can be obtained by $L_Q = \sqrt{H_B \cdot W_B}$. In this way, the area of the box remains unaltered. Afterwards, L_Q is expanded by a scaling factor s, yielding the new side length $\tilde{L}_Q = s \cdot L_Q$, such that we have larger context area as well as enough margin for data augmentation (DA). The factor is empirically set to $s = 1.8$ in this work. Finally, the cropped image is resized to match the input dimension of the CNN.

2.2 Network Architecture

Figure 2 gives a schematic overview of our effective one-pass approach. Similar to CNNs for other vision tasks, it is composed of a back-end for LR image feature extraction with three feature-extraction blocks (FEBs), i.e., a convolution block (CB) followed

by a pooling layer, and three dedicated streams for the respective rigid shape, expression and pose recovery. The following parts of this section specify the design choices in the proposed solution for LR dense face reconstruction.

2.2.1 Back-end

Since the input dimension for the back-end is just 32×32, the biggest challenge arises when adopting existing architectures to the LR domain, where the depth of the network, or equivalently the number of FEBs, is limited by the pooling layer after the CB, which enhances the spatial invariance of the CNN, but at the same time reduces the spatial resolution of the feature maps by a factor of two. For instance, the output of the prevailing VGG16 model [21] with five pooling layers will be 1×1 in our case.

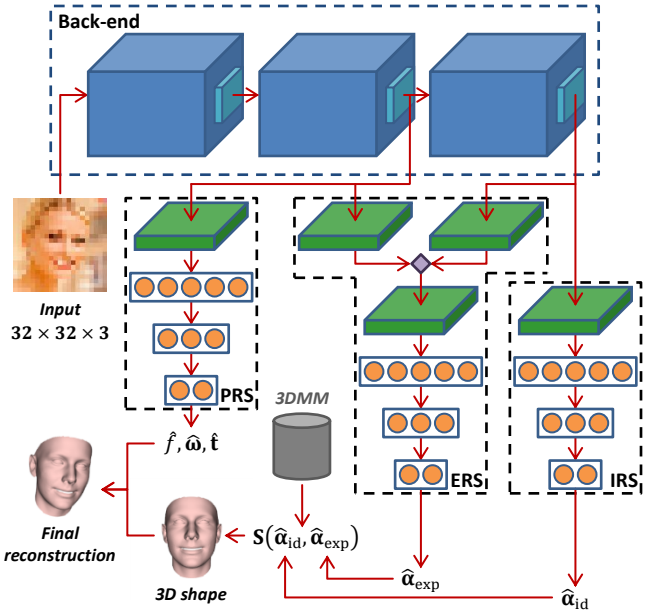


Figure 2. Overview of the LR-3DFF-Net architecture.

Therefore, the main focus in designing the back-end lies in an effective exploitation of the relatively shallow network. Inspired by the LR face recognition approach [7], which has a similar problematic nature as ours, we employ a 3-FEB architecture. Similar to VGG16, the number of feature channels after the first CB F is doubled after each additional CB, i.e., $2^{k-1}F$ for the k^{th} CB.

The CBs can be flexibly implemented with different realizations. In this work, we experiment with three possibilities, i.e., the classical and inception [22] variants in [7], and our own adaptation of VGG16 called VGG-mod, which are specified in Table 1.

The classical variant uses a single 3×3 filter in each CB to match the small content size in the LR image, which works surprisingly well for LR face recognition [7]. On the other hand, the inception module with parallel paths is considerably more complicated. The 1:4:2:1 ratio of feature maps in the respective blocks with non-uniform receptive fields accounts for multi-scale processing as in traditional vision tasks. In contrast to the classical CB, where more consecutive convolution filters seem to be counterproductive or at least unnecessary in [7], our experience regarding LR face reconstruction is exactly the opposite. In VGG-mod, we stack three convolution layers per block as in VGG16, increasing the network depth to extract more semantic information in the wake of the large parameter space in the reconstruction streams. Note the different base filter counts F for the back-ends due to their complexity, where the classical, inception and VGG-mod network has 512, 256 and 256 channels respectively.

2.2.2 Reconstruction Streams

The identity reconstruction stream (IRS), expression reconstruction stream (ERS) and pose reconstruction stream (PRS) do the essential job for our reconstruction task, which is based on the 3DMM as an intermediate medium. 3DMM is one of the seminal work for 3D face modeling, where the shape can be described as linear combinations of identity and expression parameters $\boldsymbol{\alpha}_{\text{id}}$ and $\boldsymbol{\alpha}_{\text{exp}}$. The complete 3D face is modeled through

$$\mathbf{V}(\mathbf{p}) = f \cdot \mathbf{R}(\boldsymbol{\omega}) \cdot \mathbf{S}\left(\boldsymbol{\alpha}_{\text{id}}, \boldsymbol{\alpha}_{\text{exp}}\right) + \mathbf{t}_{\text{3D}} \tag{1}$$

by the parameter vector $\mathbf{p}$, which contains in addition to the 3DMM shape coefficients the scaling factor $f \in \mathbb{R}$, 3D rotation vector $\boldsymbol{\omega} \in \mathbb{R}^3$ and 2D translation $\mathbf{t} \in \mathbb{R}^2$. By employing the 3DMM by Zhu et al. [25], we have $\boldsymbol{\alpha}_{\text{id}} \in \mathbb{R}^{199}$ and $\boldsymbol{\alpha}_{\text{exp}} \in \mathbb{R}^{29}$ respectively.

Table 1. Architecture of the k^{th} CB for three back-end types.

Classical ($F = 512$)

layer	size	stride, pad
conv	$3 \times 3 \times 2^{k-1}F$	1,1
conv	$3 \times 3 \times 2^{k-1}F$	1,1

Inception ($F = 256$)

layer	size	stride, pad
conv	$3 \times 3 \times 2^{k-1}F$	1,1
concat	$\begin{bmatrix} \begin{bmatrix} 1 \times 1 \times 2^{k-1}F/16 \\ 3 \times 3 \times 2^{k-1}F/8 \end{bmatrix} \\ \begin{bmatrix} 1 \times 1 \times 2^{k-1}F/4 \\ 3 \times 3 \times 2^{k-1}F/2 \end{bmatrix} \\ \begin{bmatrix} 1 \times 1 \times 2^{k-1}F/4 \end{bmatrix} \\ \begin{bmatrix} \text{pool } 3 \times 3 \\ 1 \times 1 \times 2^{k-1}F/8 \end{bmatrix} \end{bmatrix}$	$\begin{bmatrix} 1,0 \\ 1,1 \\ 1,1 \\ 1,0 \\ 1,1 \\ 1,0 \\ 1,1 \\ 1,0 \end{bmatrix}$

VGG-mod ($F = 256$)

layer	size	stride, pad
conv	$3 \times 3 \times 2^{k-1}F$	1,1
conv	$3 \times 3 \times 2^{k-1}F$	1,1
conv	$3 \times 3 \times 2^{k-1}F$	1,1

* Underlined layer only present in the 1st CB.

Obviously from the example input of Figure 2, it is very challenging to obtain the large parameter vector $\mathbf{p}$ for 3D geometry and pose from a single LR face image. Hence for this purpose, we build three branches to separately reconstruct the identity, expression and pose parameters for the 3DMM, i.e., IRS, ERS and PRS, which operate on different FEBs in the back-end. Note that their structures, specified in Table 2, remain unchanged no matter which back-end type is plugged.

Table 2. Architecture of the three reconstruction streams.

IRS		
layer	**size**	**stride, pad**
conv	$1 \times 1 \times 2F$	1,0
fc	4096	
fc	1024	
fc	199	
ERS		
layer	**size**	**stride, pad**
concat	$\begin{bmatrix} 3 \times 3 \times 2F \\ 1 \times 1 \times 2F \end{bmatrix}$	$\begin{bmatrix} 2,1 \\ 1,0 \end{bmatrix}$
conv	$1 \times 1 \times 2F$	1,0
fc	4096	
fc	1024	
fc	29	
PRS		
layer	**size**	**stride, pad**
conv	$3 \times 3 \times 2F$	2,1
fc	4096	
fc	1024	
fc	6	

IRS gets its input from the last FEB to maximize the semantic information through the entire deep CNN. Considering subtle expressions that need highest semantics as in IRS, which may also be lost in the lowest resolution level, the features from the middle FEB are extracted, downsampled with strided 3×3 convolution, and concatenated with those from the last FEB in the ERS. For our extra PRS as compared to [4], where the resolution of the feature maps more matters, the output from the second FEB is taken and processed in a similar fashion as in the first branch of ERS. Subsequently (and after an additional 1×1 convolution layer with $2F$ channels for each of IRS and ERS to reduce the number from $4F$), for all three streams, fully connected layers with 4,096 and 1,024 neurons are attached before the final ones with the respective target neurons.

2.2.3 Remarks

In general, BatchNorm [11] is applied exclusively for the classical and inception architecture within the back-end after each convolution layer as stated in [7], and ReLU is used after each convolution and fully connected layer throughout the entire networks, with the exception of the last output layers in the reconstruction streams.

It is worth noting although the presented approach is on the basis of the end-to-end CNN in [4], our network has a thorough redesign, adding PRS which is helpful for 3D face SR [18] or pose-aware face recognition [23]. Moreover, the proposed resolution-aware back-ends, especially the new VGG-mod variant, are absolutely paramount to break through the LR barrier for robust 3D reconstruction.

2.3 Training

The objective of training LR-3DFF-Net is to learn the mapping from a RGB image of merely 32×32 pixels to the parameter vector $\mathbf{p} = [f, \boldsymbol{\omega}^\top, \mathbf{t}^\top, \boldsymbol{\alpha}_{\text{id}}^\top, \boldsymbol{\alpha}_{\text{exp}}^\top]^\top$, which models alongside 3D geometry also the pose of the face in an end-to-end manner, contrary to the cascaded one in [25].

Because of our more challenging problem setup with extra pose estimation on input images of much lower spatial resolution compared to [4], the simple mean square error (MSE) loss on the mere 3D vertices for identity and expression, respectively, is not adequate. However, in order to prevent dominating pose error in the initial training phase, we refrain from integrating pose parameters into the MSE loss on the 3D vertices alone

$$E_{\text{shape}} = \left\| (\mathbf{A}_{\text{id}}, \mathbf{A}_{\text{exp}}) \cdot \boldsymbol{\alpha} - (\mathbf{A}_{\text{id}}, \mathbf{A}_{\text{exp}}) \cdot \hat{\boldsymbol{\alpha}} \right\|_2^2. \quad (2)$$

Here $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\text{id}}^\top, \boldsymbol{\alpha}_{\text{exp}}^\top)^\top$ and $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_{\text{id}}^\top, \hat{\boldsymbol{\alpha}}_{\text{exp}}^\top)^\top$ are the ground truth and estimated 3DMM parameters, and $\mathbf{A}_{\text{id}}$ and $\mathbf{A}_{\text{exp}}$ denote identity and expression matrices respectively.

Instead, we choose to follow [19], imposing a dedicated pose loss

$$E_{\text{pose}} = \|\mathbf{q} - \hat{\mathbf{q}}\|_2^2 \quad (3)$$

for our final loss function

$$E = E_{\text{shape}} + \lambda E_{\text{pose}}, \quad (4)$$

where $\mathbf{q} = (f, \boldsymbol{\omega}^\top, \mathbf{t}^\top)^\top$ and $\hat{\mathbf{q}} = (\hat{f}, \hat{\boldsymbol{\omega}}^\top, \hat{\mathbf{t}}^\top)^\top$ are the ground truth and estimated pose parameters, and λ is the weighting factor for balancing the two losses.

3. EVALUATION

3.1 Experimental Setup

3.1.1 Datasets and Metrics

The publicly available in-the-wild 300W–LP and AFLW2000–3D datasets are adopted, where the ground truth 3DMM parameters in 300W–LP are acquired by synthetic profiling of faces from near-frontal poses [25].

We follow the workflow in Section 2.1 to generate 32×32 input images and their 3D ground truth. Thanks to the capacity of the incorporated detector [10], 122,287/122,450 and 1,997/2,000 images from 300W–LP and AFLW2000–3D are obtained for training and evaluation respectively.

The normalized mean errors (NMEs) w.r.t. the bounding boxes for 68 2D landmarks [5] and 53,215 3D model vertices are used.

3.1.2 Implementation Details

We implement our CNN using the Caffe framework [13]. Also the face detector [10] is ported to Caffe for seamless integration into our LR pipeline. The Adam solver [16] is utilized with a learning rate of 10^{-4} on batches of 32 images. Due to different network complexities, initial training with 45k iterations for the classical back-end, and 60k for inception and VGG-mod variants is performed, before the learning rate is downscaled twice by 0.1 for two further runs of 15k iterations each. The weight in the multi-task loss is set to $\lambda = 10^{11}$ to get approximately the same initial error for both terms [19].

3.2 Results

DA is a useful tool for effortlessly enlarging data size for training deep CNNs. Nevertheless, the impact of DA is not yet clear for 3D reconstruction, as the adjustment of ground truth 3D pose

Table 3. Impact of DA on 2D (left) / 3D (right) NMEs for LR-3DFF-Net.

level	0		1		2		3		4	
rotation, shift	0°, 0%		7.5°, 5%		15°, 10%		22.5°, 15%		30°, 20%	
LR-3DFF-Net (classical)	0.0854	0.1087	0.0791	0.0983	0.0793	0.0975	0.0848	0.1026	0.0935	0.1108
LR-3DFF-Net (inception)	0.0878	0.1116	0.0811	0.1007	0.0801	0.0984	0.0839	0.1018	0.0920	0.1096
LR-3DFF-Net (VGG-mod)	**0.0802**	**0.1036**	**0.0726**	**0.0918**	**0.0702**	**0.0883**	**0.0743**	**0.0921**	**0.0770**	**0.0948**

Table 4. Comparison w.r.t. 2D (left) / 3D (right) NMEs against state-of-the-art methods on different IOD sets.

set	hard		moderate		easy		all	
IODs	{3, 4}		{5, 6}		{8, 10}		{3, 4, 5, 6, 8, 10}	
3DDFA [25]	0.1678	0.2066	0.1001	0.1295	0.0685	0.0910	0.1121	0.1424
FAN [2]	0.3493	—	0.1349	—	**0.0513**	—	0.1785	—
LR-3DFF-Net (classical)	0.0919	0.1119	0.0752	0.0932	0.0708	0.0875	0.0793	0.0975
LR-3DFF-Net (inception)	0.0918	0.1117	0.0760	0.0938	0.0727	0.0896	0.0801	0.0984
LR-3DFF-Net (VGG-mod)	**0.0816**	**0.1015**	**0.0666**	**0.0841**	0.0624	**0.0793**	**0.0702**	**0.0883**

parameters is not trivial. In this work, we experiment with in-plane rotation and translation to enrich the training data on LR faces of mixed sizes, where the extent of rotation and shift (w.r.t. the ROIs) is categorized into five levels, i.e., none, 7.5°/5%, 15°/10%, 22.5°/15% and 30°/20%. In Table 3, the benefits of DA for LR-3DFF-Net are verified. Obviously, a moderate amount of DA until the 4th level is in general favorable for all three back-end types, although the profit already saturates after the 1st stage for the shallowest classical variant, implying underfitting for more complex cases. In overall, the proposed VGG-mod achieves the lowest 2D and 3D errors throughout the evaluation, while the 2nd augmentation level appears to be the "sweet spot" for LR-3DFF-Net independent of the back-end choice.

Figure 3. Example reconstruction results on the real-world surveillance 3DPeS data [1].

In Table 4, we benchmark our LR-3DFF-Net against two state-of-the-art 3D reconstruction and 2D face alignment methods, namely 3DDFA [25] and FAN [2]. For systematical performance assessment w.r.t. different degrees of difficulty, we group faces spanning various interocular distances (IODs) into three subsets, the hard one with three to four pixels, the moderate one with five to six pixels, and the easy one with eight and ten pixels, which are then interpolated to match the respective network input sizes.

Within our LR-3DFF-Net variants, the trend is close to that of the last experiment, i.e., the classical and inception back-ends work similarly well, but both are outperformed by VGG-mod. In contrast, 3DDFA, which is targeted for HR faces, is considerably worse in LR situations than our LR-3DFF-Nets, particularly on the hard and moderate sets with just three to six pixels of IOD. The gap only becomes smaller on the easy set. Surprisingly, FAN is remarkably inferior even to 3DDFA in most scenarios except for the easy one, probably in consequence of localization ambiguity for heatmaps on very LR lattices. This outcome conforms to that in [2], where FAN performance starts to drop significantly for faces smaller than 30×30 pixels. Correspondingly, our images with an average IOD of eight pixels are approximately 36×36 pixels in size. It is noteworthy that FAN is a dedicated landmark detection approach, whereas our LR-3DFF-Net solves the more challenging dense reconstruction problem. The 2D landmarks are projected from 3D models using fixed annotations without being explicitly optimized for as in FAN, which could cause higher 2D NMEs. Based on the observations above, we can conclude that the proposed work is effective for 3D fitting on extremely LR faces.

Finally yet importantly, qualitative results on images from real-world surveillance sequences are demonstrated. The 3D face in Figure 1 is reconstructed using the LR image extracted from [14]. In spite of the strong and unknown blurring kernel, the pose, shape, and expression of the target person can be faithfully recovered. In Figure 3, the images in the 3DPeS dataset [1] show a broader range of variations, including blurring, compression artifact, pose, and even occlusion with sunglasses. Nevertheless, our LR-3DFF-Net is still capable of handling these challenges.

4. CONCLUSIONS

This paper presents a novel, pragmatic and first ever end-to-end CNN pipeline for dense face reconstruction from a single LR image, which is a result of coupling a state-of-the-art face detector, a carefully chosen back-end for resolution-aware feature extraction, and dedicated multi-task branches for recovering 3D face pose and shape models simultaneously. Extensive experiments justify the effectiveness of our LR-3DFF-Net against prior arts.

5. REFERENCES

[1] D. Baltieri, R. Vezzani, and R. Cucchiara, "3DPeS: 3D people dataset for surveillance and forensics," in *J-HGBU*, 2011.

[2] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *ICCV*, 2017.

[3] A. Bulat and G. Tzimiropoulos. (2017). Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. arXiv: 1712.02765 [cs.CV].

[4] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3D face reconstruction with deep neural networks," in *CVPR*, 2017.

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," Image *and Vision Computing*, 2010.

[6] C. Herrmann, C. Qu, and J. Beyerer, "Low-resolution video face recognition with face normalization and feature adaptation," in *ICSIPA*, 2015.

[7] C. Herrmann, D. Willersinn, and J. Beyerer, "Residual vs. inception vs. classical networks for low-resolution face recognition," in *SCIA*, 2017.

[8] G. Hu, "Face analysis using 3D morphable models," PhD thesis, University of Surrey, 2015.

[9] G. Hu, C. H. Chan, J. Kittler, and W. Christmas, "Resolution-aware 3D morphable model," in *BMVC*, 2012.

[10] P. Hu and D. Ramanan, "Finding tiny faces," in *CVPR*, 2017.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[12] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," in *ICCV*, 2017.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. (2014). Caffe: Convolutional architecture for fast feature embedding. arXiv: 1408.5093 [cs.CV].

[14] Y. Jin and C.-S. Bouganis, "Robust multi-image based blind face hallucination," in CVPR, 2015.

[15] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *CVPR*, 2016.

[16] D. P. Kingma and J. Ba. (2014). Adam: A method for stochastic optimization. arXiv: 1412.6980 [cs.CV].

[17] P. Mortazavian, J. Kittler, and W. Christmas, "3D morphable model fitting for low-resolution facial images," in *ICB*, 2012.

[18] C. Qu, C. Herrmann, E. Monari, T. Schuchert, and J. Beyerer, "Robust 3D patch-based face hallucination," in *WACV*, 2017.

[19] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *CVPR*, 2017.

[20] E. Richardson, M. Sela, and R. Kimmel, "3D face reconstruction by learning from synthetic data," in *3DV*, 2016.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[23] Z. Wang, Z. Miao, Q. M. J. Wu, Y. Wan, and Z. Tang, "Low-resolution face recognition: A review," *Visual Computer*, 2014.

[24] H. Yang, X. Jia, C. C. Loy, and P. Robinson. (2015). An empirical study of recent face alignment methods. arXiv: 1511.05049 [cs.CV].

[25] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *CVPR*, 2016.

Planetary Marching Cubes: A Marching Cubes Algorithm for Spherical Space

Zackary P. T. Sin
The Hong Kong Polytechnic University
Hunghom
Hong Kong
+852 3400 8421
csptsin@comp.polyu.edu.hk

Peter H. F. Ng
The Hong Kong Polytechnic University
Hunghom
Hong Kong
+852 2766 7248
cshfng@comp.polyu.edu.hk

ABSTRACT
There is a growing interest in digital games with user-generated content. Games with user-generated content usually involve terrain editing and marching cubes is a popular algorithm that permits a dynamic terrain. On the other hand, there is also growing interest in games with a planetary theme. Hence, a question is asked on whether can marching cubes be used to generate a planetary terrain. This study investigates how to adopt the marching cubes algorithm in a spherical space, specifically, for generating a planetary terrain. The result is the proposed planetary marching cubes, which compared to previous methods, could generate more complex terrain features while retaining smooth surfaces.

CCS Concepts
• Computing methodologies ~ Mesh models.

Keywords
Mesh Generation, Marching Cubes, Game Engineering, User-generated Content, 3D Terrain Model

1. INTRODUCTION
Classically in digital games, game content such as characters, levels and gameplay itself is served to the player in a waterfall manner. Artists create the game content, deploy to the game and the game content will be static as it is shipped to the players. Although not a novel idea, recently there is a growing interest in game content creation that involves the players. It is referred to as user-generated content as the players could use the tools provided by the game to create their own content. It is said that this is not a novel idea as, as early as 2008, the species evolving simulation game Spore by Maxis allow players to edit characters, buildings and vehicles by exchanging, scaling, moving parts and more. It is argued that it is the sandbox game Minecraft by Mojang which caused an increased interest in user content generation as many digital games involving user content eneration comes after. Terraria by Re-Logic, Project: Spark by SkyBox Labs and Dragon Quest Builder by Square Enix to name a few. One of the key user-generated content elements in these games is that it allows the player to edit the game world's terrain.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301522

Since the terrain could be modified by the players, the terrain model could not be prepared as a static model by the artists. Instead, the model needs to be supported by some algorithm to anticipate change. To generate a smooth 3D terrain that is modifiable, a popular algorithm or perhaps the algorithm to use is the marching cubes algorithm by Loresnsen [1]. In short, the idea of the algorithm is to use a 3D grid as the volume to bond the terrain. Each edge of a cube in the 3D grid consists of a parameter, or isovalue. By tuning the isovalue, the terrain mesh could be dynamically generated by computing the isosurface based on the isovalues. The marching cubes' 3D grid exists in the canonical x, y and z spatial space.

On the other hand, there also seems to be a growing interest in games based on planets. Games like Super Mario Galaxy by Nintendo, Universim by Crytivo Games and ECO by Strange Loop Games feature planets where the players could interact with. Of course, this would require a 3D model of the planet to be deployed into the game.

Here, we ask the question, how could we use marching cubes to generate a planetary terrain? This is a question worth investigating as originally marching cubes is for a canonical 3D space while a planet is best described in a spherical space. This is likely why we do not use x, y and z to describe a location on earth and instead use a geographic coordinate system to do so. Although a marching cubes algorithm could still be used to directly generate a planetary terrain, there exists a few problems. Mainly, there is an issue regarding the incompatibility of using 3D-grid-bound parameters to approximate a planetary surface. It could be imagined that the same terrain feature at different position on the planet would requires different parameters which is not optimal. Ideally, the same terrain feature regardless of positions would require similar parameters such that the terrain feature's meshes are as similar as possible. Another problem is that if we are to generate a specific planetary model and a planet is best described in spherical space, how could we, with respect to the planet, infer the suitable parameters to generate its mesh? As can be seen, directly using a typical marching cubes algorithm for generating a planetary model could cause some cumbersome problems.

In this study, planetary marching cubes (PMC), a variant of the marching cubes algorithm is proposed to generate planetary terrain instead (Figure 1). In more general terms, the proposed algorithm is a marching cubes variant that works in spherical space. To tackle

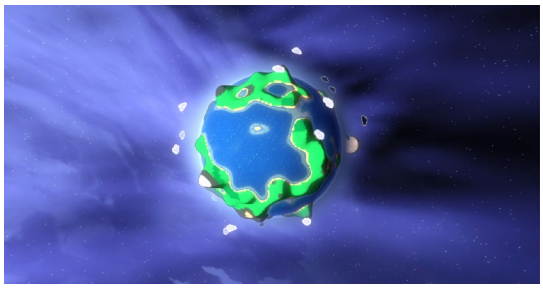


Figure 1. A planet 3D model generated with planetary marching cubes, a marching cubes algorithm adopted for generating mesh in a spherical space.

the problems aforementioned, the proposed algorithm uses a grid in spherical space instead of the canonical 3D spatial space as shown in Figure 2. So, the same terrain feature will use similar parameters to generate while it is also trivial to infer suitable parameters to generate a specific planetary terrain. In addition, it is likely desirable to generate a navigational data structure for navigation of agents played by the computer. However, since the terrain could be edited by the players, a typical navigational mesh baking method which take noticeable amount of time is likely not suitable. Hence, how a navigational grid could be conveniently generated by PMC is also proposed. Aside from generating terrains for planetary-themed games, it is expected that the proposed model could also be useful for generating 3D models for real celestial entities such as Mars. Which, could be useful for scientific education and perhaps even practical simulation.

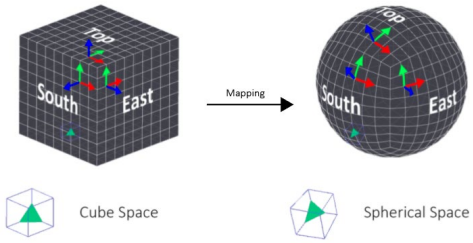


Figure 2. Projection from a unit cube to a unit sphere. The key idea of PMC is to use a spherical space instead of a canonical space for the marching cubes. Original image from [9] and modified.

2. LITERATURE REVIEW

As a computer graphics algorithm developed by Lorensen [1], marching cubes is used for generating a polygonal mesh of an isosurface by using a three-dimensional scalar field. Outside the game industry, marching cubes is usually used for medical visualization. In games, on the other hand, it is mostly used for terrain generation that could support terrain editing. There are many terrain engines that are built upon the marching cubes algorithm and the terrain of Project: Spark is likely to use marching cubes, or an adaptation, for its terrain as well.

The principal idea of the algorithm is to use a 3D grid as the volume to bond the terrain. Since it is a 3D grid, there are cubes. Each corner of a cube is what referred to as a voxel. The voxel is basically a bit that could be on (filled) or off (unfilled). Depending on the on and off combination, a cube would generate a localized mesh as shown in Figure 3. Each edge of the cube consists of an isovalue which could be tuned. By tuning the isovalues, where the mesh (isosurface) meets the edges will be changed.

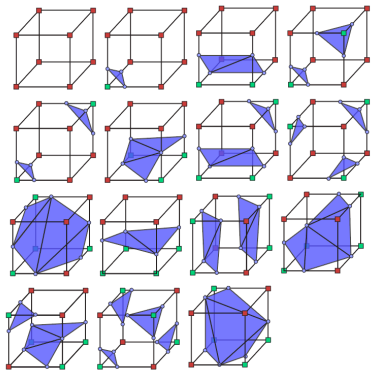


Figure 3. The 15 marching cube combinations. A small colored square represents a voxel state while a small circle represents an isovalue. Each combination will result in a different localized mesh. The blue triangles are the generated polygon.

The vanilla marching cubes algorithm or its extended algorithms could be used for generating the mesh of the planet as well, however, as mentioned, the data structure and subsequently the mesh will not be completely compatible. There will be artifacts at the eight "corners" of the planet as the volume the marching cubes covered is basically a cube with infinite size. For example, when the player is trying to create a hill feature, the hill will look significantly different at the cube's "face" and "corner". This phenomenon is basically due to the issue of unfair sampling. The axes do not align with the planet's surface and hence similar surface model will display different behavior due to different marching cube's parametrization for mesh generation. This is undesirable as at different position on the surface, the same surface model cannot be converted into the same mesh.

Adaptive methods, either adaptive marching cubes or tetrahedrizations could be used to ease the problem of unfair sampling [2, 3]. However, they could not resolve the issue of sampling as they only increase the resolution at the "corners" to attempt to reduce the difference between "corners" and "non-corners". There is also research regarding how to produce more accurate isovalues [4]. But similarly, it could only reduce the difference. It still remains true that the axes do not align with the planet's surface at the corners and as such similar model may be represented by different parameters, resulting in inhomogeneity.

Most of the previous studies on marching cubes have focused on solving ambiguity, improving performance, and extending the basic approach [5]. For example, trying to solve the face ambiguity problem that may occur in some patterns of the 256 possible marking scenarios for a cube when there are multiple facetization that could be done for the same pattern.

There exist a few marching cubes research works that are related to non-cube spaces, but they are not related to solving the sampling issue, a main focus of this study. Ref. [6] has proposed a method about reducing triangular mesh complexity (e.g. reducing the number of vertices) by reparametrizing the mesh representation in marching cubes to one represented by nearest neighbor coordinates with parameter domain in a unit sphere. Ref. [7] has proposed a

method of creating a marching cubes mesh from data collected with coordinates in cylindrical and spherical coordinates.

3. PREVIOUS METHODS

In this section, previous common methods in generating a planetary mesh are discussed. As these are methods used by the game development community, they should be easily found on the Internet.

Generally, it seems that two approaches are most popular. The first one is to use a height map on a spherical grid to generate the terrain of the planet (Figure 4). A height map is simply a 2D grid with heights representing as values. When generating the terrain, each height value will be used to infer how high the vertex should be. The planet will be divided into six faces, and each face will be represented by a heightmap. Although using height maps could create a smooth surface, it also means that more complex terrain features such as a cave or overhangs could not be represented. The game Universim seems to use this method as its terrain mesh does not exhibit complex features such as caves or overhangs.

Figure 4. A planet generated using a heightmap by Romain (https://github.com/Roldak/PlanetGeneration).

The second method is to use a voxel engine on a sphere to generate the terrain of the planet (Figure 5). Voxel engine refers to an idea of using voxels to approximate a 3D model. Not to be confused with voxel from marching cubes, each voxel in a voxel engine is simply a 3D block mesh that either exists or not. The engine will choose the relevant voxels to exist to approximate the 3D model. Games like Minecraft uses a voxel engine to generate their terrains. Similar to using heightmaps and PMC, the planet will be firstly divided into six faces. Then, each face will have its own set of voxels for generating the terrain. The result is understandably "blockish" as a voxel engine could only generate terrain mesh by choosing the correct set of voxels.

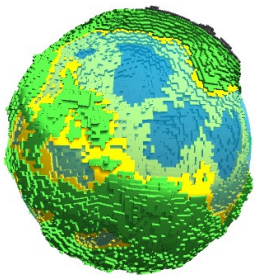


Figure 5. A planet generated using a spherical voxel engine by Frankson (https://forum.unity.com/threads/spherical-voxel-engine-cc-by.259121/#post-1721678).

4. METHODOLOGY

Here, the technical details of PMC are presented. Several problems need to be addressed. These includes how to create of a spherical space, how to convert between the coordinate system between the faces, how to organize the object hierarchy and how to resolve the inconsistency at the border of faces. In addition, how to generate a navigational grid and how to procedurally generate the terrain will also be presented.

4.1 Six Faces Voxel System

To define the spherical space of PMC, a common cube-sphere mapping equation is used:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} x\sqrt{1-\frac{y^2}{2}-\frac{z^2}{2}+\frac{y^2z^2}{3}} \\ y\sqrt{1-\frac{z^2}{2}-\frac{x^2}{2}+\frac{z^2x^2}{3}} \\ z\sqrt{1-\frac{x^2}{2}-\frac{y^2}{2}+\frac{x^2y^2}{3}} \end{bmatrix} \quad (1)$$

Eq. (1) will convert a unit cube coordinate (x, y, z) to a unit sphere coordinate (x', y', z'). So, (1, 1, 1) will be mapped to $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$.

This equation is quite known in the game development community. The previous methods in section 3 very likely also used this equation to create a spherical space. By converting necessary unit cube coordinate to a unit sphere coordinate, naturally, the sphere will be divided into six faces just like the cube. The base squares of the six faces will be subsequently projected on the related spherical space as well. Fig. 2 shows the result of the projection. For the height map and voxel engine approach mentioned previously, the necessary coordinates are created similarly in this step. Here for planetary marching cubes, the six faces on the unit sphere are also used to create six 3D grids. Each grid G will have its own set of voxels v.

It is proposed that each of the faces of the sphere should act as its own localized space. As shown in Figure 6, each local face should have its own coordinate system. In the Figure, only the two axes are shown. There should be one more axis, which is called the "depth" axis, that point outwards from the sphere center, hence, a localized three-axis system is formed.

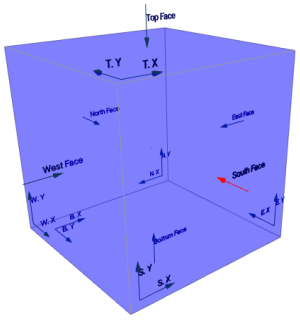


Figure 6. The six local face coordinate systems.

It is suggested that the six faces should be referred with the following names F_{Top}, F_{Bottom}, F_{North}, F_{South}, F_{East}, and F_{West}. The convention suggested for a sphere is that there is no rotation and the camera is looking the sphere from the "back" of the world space. F_{Top} and F_{Bottom} are the faces at the top and bottom of the "screen" respectively; F_{North} and F_{South} are the faces further and

closer respectively; and F_{East} and F_{West} are the faces at the left and right respectively.

4.2 Face Coordinate Conversion

There will be many instances, related to data structure or game logic, where the conversion between the six localized face coordinates become necessary. For example, if a character has move far enough and reached the boundary of a face. Although seemingly a basic operation, the conversion is not entirely trivial as neighboring faces will have a varying degree of differences. In Figure 7, we can observe that the F_{Top} and F_{South} have directly compatible face coordinates. So, the conversion will indeed be a trivial addition operation. However, the coordinates between F_{Top} and F_{East} not only do not have their x and y axes aligned, but also have one pair of their counterpart axes in the opposite direction

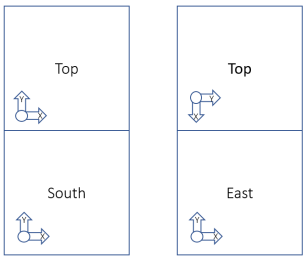


Figure 7. Neighboring local face axis systems have different degree of incompatibility. The left showcases an easy conversion, while the right showcases a problematic one as the two-axis systems are no aligned.

To this end, a coordinate conversion algorithm that could be used to convert coordinate from one face to another using simple addition, multiplication and flipping operations is devised. See **Algorithm 1** for converting the coordinates on one axis to another.

Algorithm 1 Pseudocode of the face coordinates conversion algorithm.

```
Input: Two faces, f_c, the current face, and f_t, the target face, one axis a_c and one 1D position p_c.
Output: One 1D position p_t.
Let a_t as the target axis
if f_t axis a_n with the same name as a_c is orthogonal to a_c
    a_t := OrthogonalAxis(f_t, a_n)
else
    a_t := a_n
end if
var mul := 1
var offset := 0
if IsOppositeDirection(a_t, a_c)
    offset := negate of max value of a_c
    mul := -1
else
    if position of origin of a_t has a difference with that of a_c can be measure by either a_t or a_c
    offset := OriginDifference(a_t, a_c)
    end if
end if
p_t := p_c * mul + offset
```

In **Algorithm 1**, OrthogoalAxis() retrieves the other non-depth axis of the local face axis system; IsOppositeDirection() checks if two axes are pointing different direction; and OriginDifference() retrieves the 1D difference between the origins using the axes (e.g. OriginDifference(y_{top}, y_{south}) is the size of the face while OriginDifference(x_{top}, x_{south}) is 0).

4.3 Engine Hierarchy and Parameters

In our implementation, a similar architecture to a typical voxel engine is used. As shown in Figure 8, A hierarchy of planet, faces, chunks, voxels and voxels' edges is suggested (with the former being the parent of the latter). As stated, a planet will have six faces, with each face having its own chunks. At the base of a face, there should be n_c^2 chunks as the chunks will fill the square planet face with each side having n_c chunks. For each chunk, it will have n_t^3 tiles with the width, height and depth having n_t tiles. A tile is simply a conceptual construct made up of eight voxels. It is basically the marching cube, but in this instance, the cube is not a strict cube. It has been morphed by the spherical space into a trapezoidal prism that has a smaller bottom and a larger top as shown in Figure 9. Each voxel will have three voxel edge that each expands along one of the three axes of the localized face axis system. Since each voxel edge will connect the nearby voxel, aside from the voxels at the border of a face, each voxel will be connecting to six voxel edges. Voxel will have a state of on and off to determine if it is dense and the voxel edge could be considered as an objectification of marching cubes' isovalue. Unlike the other two types of voxel edge, the type of voxel edge that expands along the depth axis has an undefined length l_d. Additionally, a base depth d_b should be set to define the radius of the unit sphere where the six faces are built. Setting a larger d_b could save memory on needless data at the core of the planet. d_c should be defined for specifying the number of chunks along the depth axis as, like l_d, it is undefined but d_c could be infinite. If d_c is infinite, chunks should be loaded dynamically

$$\eta_c = 6n_c^2 \cdot d_c \tag{2}$$

$$\eta_v = 6 \cdot (n_t \cdot n_c + 1)^2 \cdot (n_t \cdot d_c + 1) \tag{3}$$

$$r = d_b + l_d \cdot (n_t \cdot d_c) \tag{4}$$

In summary, for the planet, the total number of chunks η_c, voxels η_v and its radius r will be (2), (3), and (4) respectively. The parameters could be tuned to change the properties of the planet.

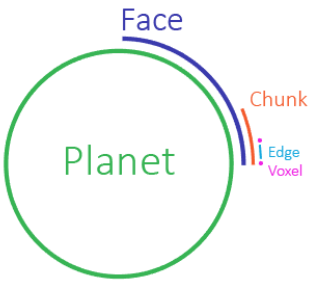


Figure 8. The hierarchy of PMC. A planet is a parent to faces; A face is a parent to chunks; A chunk is a parent to voxels; and a voxel is a parent to edges.

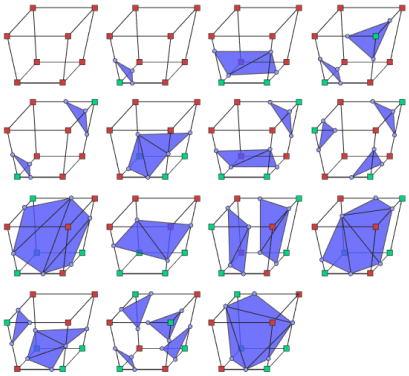


Figure 9. The 15 marching tiles combinations used in PMC.

4.4 Pairing Voxels and Edges

Unlike a simple voxel engine where each block has little connection with its neighbor, marching cubes algorithm has its voxels and subsequently voxels' edges related to their neighbor. A voxel is connected by six edges and inversely, an edge is connected by two voxels. This property will be problematic along the borders of the six faces as each face has its own local space.

For a marching cubes algorithm in canonical space, it is typical that neigboing chunks are connected by adding edges between the chunks. This solution is likely difficult and unintuitive to implement in spherical space since, unlike in a canonical space, when in a spherical space we need to consider connecting chunks internally in a face and externally between faces.

Instead of adding new edges, this paper proposes that neighboring chunks should be connected to one another by overlapping. Two chunks could be connected to each other by overlapping each other's bordering voxels. On the same face, voxels that are repeatedly overlapped should be removed such that chunks will share the bordering voxels and relevant edges. This rule, however, should not be applied for those on the faces' borders. The reasoning is that, although they may overlap, edges might be representing different directions. Ergo, our approach pairs those overlapping voxels and their edges on the border instead. By pairing the voxels and edges at the border of faces, their values could be updated to their counterparts to synchronize each other.

4.5 Navigational Grid Generation

A navigational grid is useful for navigational algorithm such as A*. Here, it is proposed that it could done via checking for terrain features such as cliffs, overhangs, and caves. When the surface is deemed "accessible", the grid could be extended to there (Figure 10).

Before considering those "inaccessible" terrain features, there is a need to check for voxels that represent a surface of the planet. Those voxels that are on the surface of the chunk's mesh are referred to as surface voxels. A surface voxel could be found by checking whether the voxel above it is filled or not. If it is not filled, the voxel is considered as a surface voxel. That is a voxel $v^{(x,\,y,\,h)}$ is considered a surface voxel when $v^{(x,\,y,\,h)} = 1$ and $v^{(x,\,y,\,h+1)} = 0$, where $(x,\,y)$ denotes the local face spatial index, h denotes the height index, v is 1 when filled and v is 0 when unfilled.

It is observed, however, that the navigational grid is not simply a collection of surface voxels. With the example in Figure 11 as a reference, it can be seen that overhangs and caves could block a path if nearby surface voxels are directly used to determine the surface area. Agents would be seen passing through the graphical mesh. Instead, it is proposed that when a surface voxel is searching for true nearby surface voxels, meaning surface voxels that could be "accessible", the algorithm should also check for those overhangs and caves. In parallel with finding the nearby surface voxels, an overhang could be found by checking if any voxel above is suddenly filled. The checking should be done iteratively, voxel by voxel, until the true surface voxel nearby is found. A cave could similarly be found as a cave is simply an extended overhang.

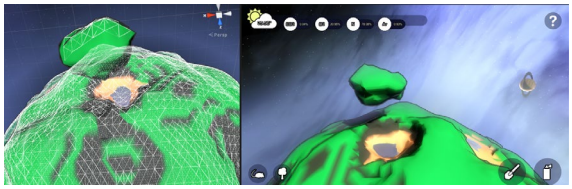


Figure 10. The result of the proposed navigational grid generating method. It could create a navigating grid when the surface is accessible and detach grids when they cannot reach each other.

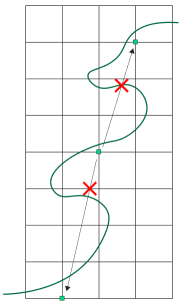


Figure 11. An illustration of an overhang, or cave blocking the path to neigbouring surface voxels at the red crosses. They are considered "inaccessible". The green line represents the terrain surface while the grid is the collection of tiles in PMC.

4.6 Terrain Procedural Generation

For a marching cubes algorithm implemented with canonical 3D axes, without the PMC algorithm, applying Perlin noise [8] to generate a planetary terrain would not be intuitive.

On the other hand, with PMC, procedural terrain generation is possible by using Perlin noise. The first intuitive method of generating the terrain could be using a 2D Perlin noise to generate a terrain for each face. Hence, there will be six different 2D Perlin noise sampling. But this requires tedious smoothing between the faces. This paper instead proposes to use fractal 3D or higher dimensions Perlin noise to generate the terrain (Fig. 12). The most simple and intuitive example is to generate the height map for the sphere, and hence the planet, using 3D Perlin noise.

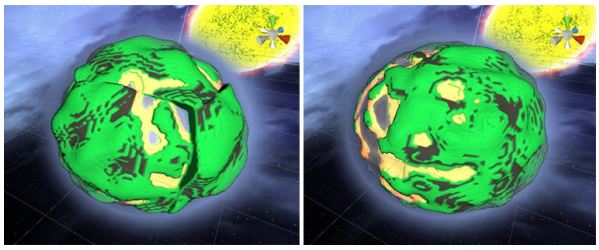


Figure 12. The difference between using 2D and 3D Perlin noise. Using 2D Perlin noise for generating terrain requires additional smoothing operation (Left). Using 3D Perlin noise to generate could bypass this need (Right).

It is worth noting that the scalar values of edges could be used for other algorithms to extract useful information for their own purposes. For example, the depth of the sea could be extracted by evaluating the scalar values of edges and the states of the voxel. This could be used for shading the shore of a sea.

5. RESULT AND DISCUSSION

Previous methods either uses heightmaps or voxels engines to generate planetary terrain. The disadvantages of the two methods are that the former could not express complex terrain features such as caves or overhangs while the latter could only generate the terrain in blocks. With PMC, it is possible to generate a planetary terrain that could express those complex terrain features while maintaining smooth surfaces as shown in Figure 13. Although the terrain allows dynamic editing, a navigational grid could also be dynamically generated in parallel (Figure 14). Lastly, 3D Perlin noise is proposed to procedurally generate the terrain. The usefulness of PMC in STEM education is shown in a previous study [9]. Therefore, it is expected that PMC could be useful for practical application such as scientific education and simulation aside from digital games.

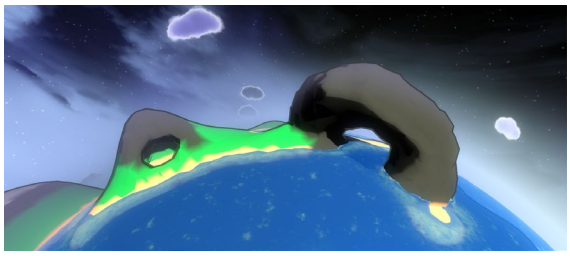


Figure 13. This figure illustrates that PMC permits the generation of complex terrain features such as caves and overhangs.

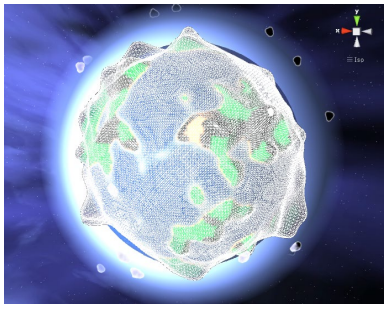


Figure 14. This figure shows the navigational grid of a PMC-generated planet.

A limitation in this research is that the PMC-driven mesh will continue to be coarser as the planet gets further from its core. A possible solution to ease this elevation-related distortion could be done by having a different n_t at different level. The new n_t should be set such that it is divisible by the previous one so that the previous tile could entirely cover the new tiles in integral number. But this solution will not truly resolve this distortion and seems to be unavoidable for spherical parametrization.

6. CONCLUSION

To generate more complex and smooth terrain features not achievable by heightmaps and voxel engines for planets, PMC, a marching cubes algorithm for spherical space is proposed. Some challenges such as the coordinate systems, navigational grid generation and procedural terrain generation have been explored.

7. REFERENCES

[1] W. E. Lorensen and H. E. Cline, "Marching Cubes: A high resolution 3D surface construction algorithm," *ACM Siggraph Computer Graphics,* vol. 21, no. 4, pp. 163-169, 1987. DOI= https://doi.org/10.1145/37402.37422

[2] R. Shu, C. Zhou and M. S. Kankanhalli, "Adaptive Marching Cubes," *The Visual Computer,* vol. 11, no. 4, pp. 202-217, 1995. DOI= https://doi.org/10.1007/BF01901516

[3] H. Muller and M. Wehle, "Visualization of Implicit Surfaces Using Adaptive Tetrahedrizations," in *Scientific Visualization Conference*, 1997.

[4] S. Fuhrmann, M. Kazhdan and M. Goesele, "Accurate Isosurface Interpolation with Hermite Data," in *International Conference on 3D Vision*, 2015. DOI=https://doi.org/10.1109/3DV.2015.36

[5] T. S. Newman and H. Yi, "A Survey of the Marching Cubes Algorithm," *Computer & Graphics,* vol. 30, no. 5, pp. 854-879, 2006. DOI= https://doi.org/10.1016/j.cag.2006.07.021

[6] G. M. Nielson, L.-Y. Zhang, K. Lee and A. Huang, "Spherical Parameterization of Marching Cubes Isosurfaces Based Upon Nearest Neighbor Coordinates," *Journal of Computer Science and Technology,* vol. 24, no. 1, pp. 30-38, 2009. DOI= https://doi.org/10.1007/s11390-009-9201-z

[7] J. Goldsmith and A. S. Jacobson, "Marching Cubes in Cylindrical and Spherical Coordinates," *Journal of Graphics Tools,* vol. 1, no. 1, pp. 21-31, 1996. DOI= https://doi.org/10.1080/10867651.1996.10487453

[8] K. Perlin, "An Image Synthesizer," *ACM Siggraph Computer Graphics,* vol. 19, no. 3, pp. 287-296, 1985. DOI= https://doi.org/10.1145/325334.325247

[9] Z. P. T. Sin, P. H. F. Ng, S. C. K. Shiu and F.-L Chung, "Planetary marching cubes for STEM sandbox game-based learning: Enhancing student interest and performance with simulation realism planet simulating sandbox," in IEEE Global Engineering Education Conference, 2017. DOI= https://doi.org/ 10.1109/EDUCON.2017.7943069

Similarity Analysis of 3D Models Based on Convolutional Neural Networks with Threshold

Shengwei Qin
South China Institute of Software Engineering, Guangzhou University
Guangzhou, China
+86 20 87818083, 510990
qsw.sise@gmail.com

Zhong Li
School of Science, Zhejiang Sci-Tech University
Hangzhou, China
+86 20 87818083, 310018
lizhong@zstu.edu.cn

Zihao Chen
South China Institute of Software Engineering, Guangzhou University
Guangzhou, China
+86 20 87818083, 510990
czh1622@scse.com.cn

ABSTRACT
The structure of a three-dimensional (3D) model will mutate into various shapes when it does deformation. Indeed, various shapes look different from the original 3D model. What is more, the two-dimensional (2D) image of the deformed 3D model is dissimilar to the 2D image of the undeformed model. Therefore, how to find a proper method to analyze the 3D model similarity always be the research hotspot. In these years, with the wide spread of deep learning technology, the research of similarity and retrieval system of 3D models has set off a new technical revolution. However, the data processing method of the 3D model is distinct with the methods of the 2D image, which is more complex. In consequence, the paper presents the similarity analysis method of 3D models based on convolutional neural networks with threshold (t-CNN). The trained and test datasets are multi-view colored 2D images of 3D models which the color is computed by the heat kernel signature (HKS). Meanwhile, in order to improve the accuracy of retrieval, a threshold is added to the CNN before confirming the final category. The experiments show that the colored dataset construction method proposed in the paper makes the data processing easier and improves the classification precision. Similarity analysis of 3D models based on the t-CNN in recall and precision is better than the other methods.

CCS Concepts
• Computing methodologies → Computer graphics → Shape modeling → Shape analysis; Information systems → Information retrieval → Retrieval models and ranking → Similarity measures.

Keywords
Three-dimensional Model; Similarity Analysis; t-CNN; Colored dataset; Heat Kernel Signature;

1. INTRODUCTION
In recent years, with wide application of deep learning technology, more and more researchers have put it into the graphic image processing. A challenging problem in computer vision is how to acquire features of the 3D models from a 2D image. The advantage of deep learning technology is that it can perfectly solve the problem of feature extraction and classification of 2D images. Accordingly, how to realize the feature extraction and classification of 3D models combining the deep learning technology becomes a research hot point for many researchers.

At the beginning of proposed the deep learning technology, the neural network cannot directly use 3D model data as input, which use the 2D images as input in the network to analyze the similarity of 3D models. In recent years, many papers have been able to input the 3D model data into the network directly for training and testing. However, the structure of this network is much more complicated than the initial network using the 2D images as input. And it is not intuitive and easily understandable in network training.

So far, the combination of deep learning technology and 3D models for feature learning can be roughly divided into the following three types:

The first one is that the features of artificial extraction are used for network learning. First, the 3D models are reprocessed by a specific algorithm, such as HKS, SOG, etc., and some parts of features of the 3D models are extracted. These features are input into the network for re-learning and the more abstract features are obtained. For example, Xie, Z., Xu, K., Liu, L., et al. [1] proposed the Extreme Learning Machine method for 3D model shape segmentation and labeling. It manually calibrated a set of data sets with labeled which was utilized to train an ELM classifier generating the initial segmentation. The final shape segmentation was obtained by computing the smooth segmentation through a graph-cut optimization based on the initial segmentation. Since the artificially defined features do not fully represent the global and local features of the 3D models and different definitions will lead to large uncertainty in the results. Simultaneously, the limitations of artificial feature extraction also lead to few excellent articles for 3D model retrieval appearing.

The second one is that the 3D model data is changed into the Euclidean space which the converted data structure can be directly processed by the deep learning network. This data processing method has changed ontology feature form of the 3D models. Therefore, this is the most studied problem on 3D model retrieval. For instance, Su, H., Maji, S., Kalogerakis, E., et al. [2] tried to solve the 3D models retrieval problem from the 2D images. They firstly computed some multi-view 2D images from the different perspectives without processing the 3D models. And then put these 2D images into a standard CNN network for training the classifier. The experiments showed that the 3D models could be

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301529

identified from a single view image. When provided multi-view 2D images, the recognition rate would be further increasing. Wang, P. S., Liu, Y., Guo, Y. X., et al. [3] proposed an octree-based convolutional neural network (CNN) for 3D model shape analysis. The authors used the octree to represent the 3D models and sampling the normal vectors from the surface of the 3D models as input of the 3D CNN. Since 3D CNN processed the normal of the surface of the 3D models, the O-CNN method was equally applicable to the high-resolution 3D models. Nevertheless, it could not distinguish the deformed models and undeformed models. Charles, R. Q., Su, H., Mo, K., et al. [4] proposed a PointNet for the 3D point cloud model which was no easier to handle than the 3D surface model. And it was the first deep neural network for processing the out-of-order point cloud data directly. A function was designed in the PointNet, which the values of the function were independent of the order of the input data. These functions were named as symmetric functions. The authors utilized the max pooling layer as the main symmetry function. Although this method was simple, the experiment proved that the effect was better. The third one is different from the second method. The data of the 3D model in this method are no longer converted into the Euclidean space, and the data structure is directly defined by the two-dimensional manifold or graph. The typical networks are Geodesic CNN (GCNN). For example, Masci, J., Boscaini, D., Bronstein, M. M, et al. [5] proposed a geodesic convolutional neural network (GCNN) on the Riemannian manifold based on the local geodesic coordinate system. The paper used the GCNN to learn the invariant shape features of 3D models, which achieved a good performance in shape description and retrieval.

Through the above research work, the current similarity analysis of 3D models has not found a unified method. It still stays on how to process the data of the 3D models. The above research work does not show a great advantage in the retrieval. Either processing data are too troublesome or the retrieval accuracy is poor. Our paper considers how to make data processing easier and improve the retrieval accuracy for the 3D models. On the basis of the two problems, firstly, for the 2D images, the CNN network can learn its features more quickly, and the accuracy is very high. Consequently, how to describe the global and local features of the 3D model from the 2D images has discussed in the paper. Secondly, CNN is generally used for image classification. During the test, no trained images will also be classified into a similar category. Therefore, how to exclude the non-training images during the classification is considered. Meanwhile, this consideration also guarantees the classification accuracy. Consequently, this paper proposes a colored datasets construction algorithm combining with the characters of the shape of the 3D models. On the one hand, the color retains the local features of the 3D models, and color features are preserved regardless of whether the model is deformed. On the other hand, the acquisition of multi-view 2D images retains the global features of the 3D models and provides a guarantee for training and testing on CNN with threshold.

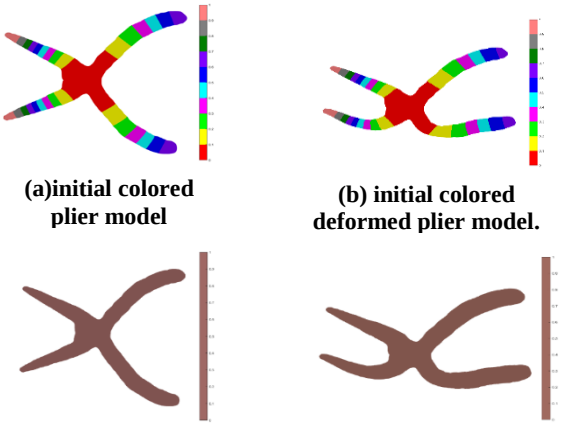
2. COLORED DATASETS

2D images are easier and convenient input types for the CNN. And the classification accuracy of 2D images is very high [6-7]. However, unlike the 2D images, the data structure of the 3D models is more complicated which the data structure is nonlinear and the data volume is larger. So how to cope with data of 3D models more easily is discussed in this section and it is more desirable to completely retain the characteristics of shape of the 3D models when converts 3D data into 2D data. Therefore, the paper will start from the following two principles to construct a dataset.

A. How to describe the global and local geometric characteristics of 3D models without extract artificial features from 3D models as much as possible.

B. How to ensure the description features still valid when the model is deformed.

Starting from the above two principles, the paper considers the validity of the features after the model deformation. Thus the paper constructs a multi-view colored datasets which will retain the global and local geometric features of the 3D models.



(a)initial colored plier model **(b) initial colored deformed plier model.**

(c) colored plier model. **(d) colored deformed plier model**

Figure 1. The colored state of plier model

Firstly, the 3D model is colored by the HKS method [8], as shown in Figure 1(a-b). The initial colored model is colored again according to the formula (1) which is for reducing the influence of color blocks on feature extraction [9]. Because the center color of the initial colored model is always red. Hence, for any models, those with a similar color distribution will always be considered similar. The final colored models are as shown in Figure 1(c-d). Even if the plier models are deformed, the color characteristics will remain unchanged.

$$\mathrm{RGB}_{final} = \frac{\sum_i \frac{A_i}{C_{total}} * RGB_i}{10} \tag{1}$$

Where A_i is the sum of the number of points where the current color is grouped as i. C_{total} is the sum of the points of the 3D model. RGB_i is the RGB value when the current color group is i $(i = 1 \ldots\ldots n)$ and n is the total number of colors of the initial colored model.

Secondly, the 3D model is placed under the different viewpoints. In order to retain all shape features of the 3D model, the paper proposes an algorithm for construction the datasets based on the viewpoint trackball, as shown in Figure 2.

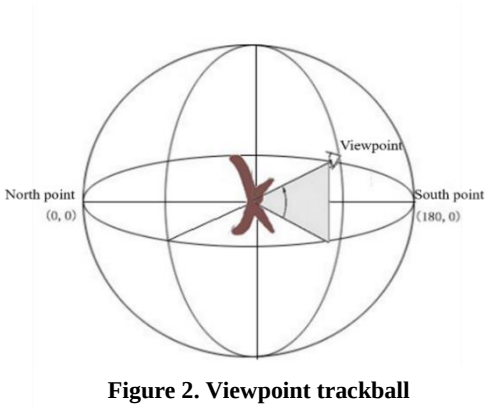


Figure 2. Viewpoint trackball

The algorithm steps are as follows:

A. Calculate the size of the AABB bounding box of the colored 3D model, and mark the length, width and height as *a*, *b*, *h* respectively. The maximum value of length of the bounding box on the x-axis is denoted as b_{xmax}, and the width of the bounding box on the y-axis is denoted as b_{ymax}. Construct a viewpoint trackball to ensure that the diameter of the trackball is the diagonal length of the bounding box plus a minimum value ε, which ensure that the 2D data acquired from different perspectives is basically full of images and it does not exceed the scope of the image.

$$\mathrm{d} = \sqrt{a^2 + b^2 + h^2} + \varepsilon, \quad \varepsilon > 0$$

B. Put the viewpoint into an arbitrary position (α_z, β_e) of the trackball, which means that you will see the center 3D model $G_i (G_i \subset G^*, G^*$is a model set) from the current position (α_z, β_e), where α_z is the azimuth, $\alpha_z \in (-360°, 360°)$ and β_e is elevation angle, $\beta_e \in (-360°, 360°)$. The starting position of the viewpoint is$(0°, 0°)$.

C. Calculate the position (x', y', z') of the viewpoint from the viewpoint angle (α_z, β_e) in the coordinate system of 3D models. Convert the coordinate system of the 3D models into a new coordinate system which the viewpoint is the origin point. Finally, map the point $(x_{G_i}, y_{G_i}, z_{G_i})$ of the 3D model in the new coordinate system to a 2D image $(x_j, y_j, z_j)_{G_i}, j = 1 \ldots |\alpha_z| \times |\beta_e|$ by the perspective projection.

$$x_j = \frac{x_{G_i} * n}{z_{G_i}}, y_j = \frac{y_{G_i} * n}{z_{G_i}}, z_j = n \tag{2}$$

Where *n* is the distance from the near clipping plane to the position of the viewpoint(x', y', z').

D. Set the step size ξ for changing the angle of viewpoint. The smaller the step size ξ, the little change of the angle of viewpoint (α_z, β_e). Meanwhile, the saved 2D images will have a little change and the datasets will be large; on the contrary, the degree of change of the saved 2D images is abruptly change which will result in loss of valid features of the 3D models. Accordingly, the change parameter ξ of the viewpoint is computed as follows:

$$\lceil \xi \rceil = \begin{cases} \frac{360^\circ}{b_{xmax}}, & b_{xmax} > b_{ymax} \\ \frac{360^\circ}{b_{ymax}}, & b_{xmax} \le b_{ymax} \end{cases} \tag{3}$$

In the viewpoint trackball, we can see that the images are the same when the viewpoints at (0, 0), (360, 360) and so on. In consequence, update the range of the viewpoint (α_z, β_e):

$$\alpha_z, \beta_e \in (-360° + \lceil \xi \rceil, 360° - \lceil \xi \rceil)$$

Repeat the above steps B to C until all perspective images have been acquired.

Finally, a colored dataset is constructed according to the above algorithm, which the 3D models are selected from McGill model set [10]. The step of changing the angle of viewpoint is 10, as shown in the Figure 3.

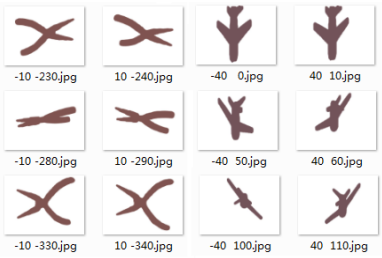


Figure 3. Colored dataset

3. t-CNN CONSTRUTION

In the section 2, the colored dataset has been constructed using the viewpoint trackball algorithm, as shown in Figure 3. This section will discuss the CNN training and verification. First, traditional CNN is constructed which can achieve the classification of the two models. The colored dataset constructed in the section 2 will be the training dataset and the size of the colored dataset is 10082. Next, a plier and airplane model will be selected from the McGill model set as the test model, which they are different from the models in the figure 3. Then, the test colored dataset of the 3D models in Figure 4 is obtained through the viewpoint trackball construction algorithm, where the step of changing the angle of viewpoint is approximately equal to 11, as shown in Figure 5. Finally, put test images into trained CNN for models classification.

Figure 4. Uncolored 3D models for testing

Figure 5. Test colored dataset.

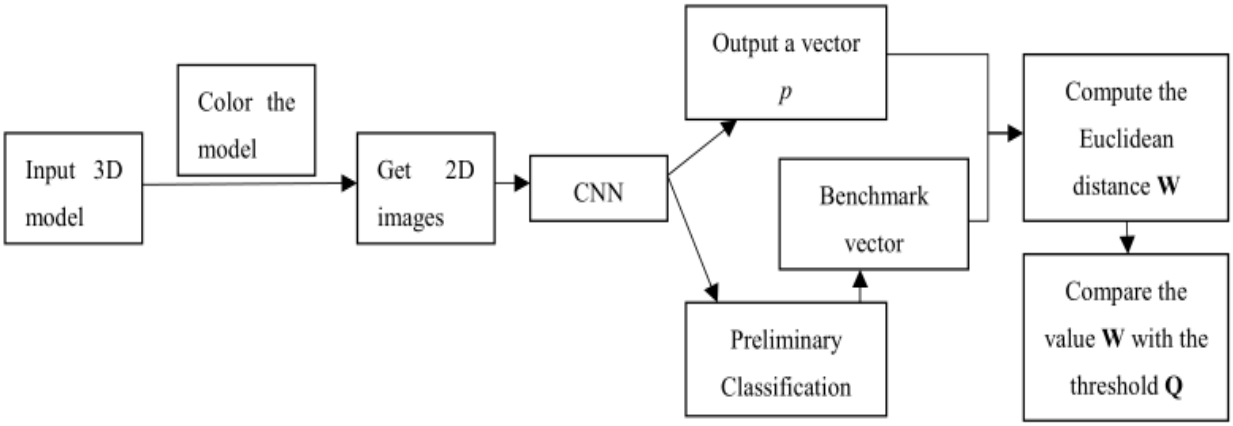


Figure 6. The flow chart for classification of 3D models.

After the experiment, the error rate of image classification is 5.2%, which some view images of the airplane models are divided into plier's category and all images of the plier model are classified correctly. Therefore, the plier model in Figure 4 is deemed to be the same model in Figures 3. Some view images of the airplane models are identified as plier model, such as the picture "-385-352.jpg" are misclassification. Is this model still an airplane model? It is not known through this experiment. So how to finally confirm the category of the model, it is a problem. Another point is that the classification of the CNN is limited, such as the color of images of the two models in Figure 5 is different from it in Figure 3. When the untrained models are put into the CNN, due to the limitation of classification of the trained models, the untrained models will always be identified as the most similar categories. Is this test plier model in Figure 4 a trained plier model in Figure 3? This is a problem for the finite classification network. Based on the two problems, the paper proposes a CNN with a threshold. For different application backgrounds, we can always find a finite classification network which will need to classify the models exactly.

3.1 CNN with Threshold

On the basis of the above problems, the process of model similarity analysis is re-optimized combined with the idea of the Siamese neural network. The result is no longer directly confirmed by the traditional CNN. We propose a CNN with a threshold, which the category of the 3D model is finally confirmed by the threshold and the threshold is obtained by the CNN. The following steps are used to compute the benchmark vector and threshold.

A. Select *v* images $(\alpha_{z_i}, \beta_{e_i}), i = 1 \dots v$ from the training colored dataset **I** of a 3D model, where α_z represents the azimuth and β_e represents the elevation angle. Selected *v* images *i*, *j* must satisfy the following conditions:

$$\left||\alpha_{z_i}| - |\alpha_{z_j}|\right| > \frac{360}{\lceil\xi\rceil} - 1, \left||\beta_{e_i}| - |\beta_{e_j}|\right| > \frac{360}{\lceil\xi\rceil} - 1$$

When α_{z_i} is determined, the value $\left||\alpha_{z_i}| - |\beta_{e_i}|\right|$ is maximum and $\alpha_z, \beta_e \in (-360° + \lceil\xi\rceil, 360° - \lceil\xi\rceil)$.

B. Put *v* images into the trained CNN and obtain an output vector $p_i, i = 1 \dots v$, which is used as the benchmark vectors of *v* views for this category of the 3D model.

C. The all images in the colored datasets **I** except the *v* images will be input into the trained CNN, which obtain *m* output vectors $p'_m, m = 1 \dots |\mathrm{I}| - \mathrm{v}$. Compute the Euclidean distance $W_{m,i}$ between the output vector p'_m and the benchmark vector p_i.

D. Compute the minimum value of *v* values of $W_{m,i}$, denoted by Q_m. The mean value of Q_m will be as the threshold Q of this category of the 3D model.

$$Q = \frac{\sum_m \min_i \mathrm{Q}_m}{|\mathrm{I}| - v}, \mathrm{Q}_m = W_{m,i}, m = 1 \dots |\mathrm{I}| - v, i = 1 \dots v.$$

Through the above steps, the *v* benchmark vectors p_i and the threshold Q of the category of the 3D model can be obtained. In the test or verification phase, when a 3D model is input, $l(l \geq 1)$ 2D images are acquired under the viewpoints(α_z, β_e). The test process is as shown in Figure 6.

A. Put the $l(l \geq 1)$ 2D images into the trained network CNN and obtain *l* output vectors $p_l(l \geq 1)$.

B. Confirm the category of *l* images of a 3D model. If all the *l* images of a 3D model are divided into the same category, go to step C. If *l* images are separated into different categories, go to step D.

C. Compute the Euclidean distance $w_{l,i}$ between the output vector p_l of *l* images and the corresponding category benchmark vectors $p_i(i = 1 \dots v)$. Since each image has *v* Euclidean distances, for convenience, compute the mean value of *v* Euclidean distances of per image, denoted by $\overline{w_l}$. And then obtain the average value W from the values $\overline{w_l}$ of *l* images. If W < Q, it is the same category with the trained model, otherwise it is considered to be dissimilar with the trained model.

D. Suppose that the *l* images are identified as $m(m \geq 2)$ categories. Compute the Euclidean distance $w_{j,l,k}$ between the output vectors p_l of *l* images and the corresponding benchmark vectors $p_{j,k}(j = 1..m, k = 1 \dots v)$ with the category. The average distance w_j is calculated and compare it with the threshold Q_j of the corresponding category *j*. If $w_j < Q_j$, it is determined to be similar with the category of a trained model. If the value w_j is smaller than the thresholds Q_j of different categories, the final

category will be determined according to the minimum Euclidean distance w_j.

3.2 t-CNN verification

In order to verify the validity of t-CNN, this section selects three different kinds of 3D models for verification, as shown in Figure 7. During the test, only one image of each model is chosen as the input of CNN. Therefore, there is no case where the part of images is divided into multiple categories. The following section will show it.

(a)bird model **(b)deformed plier model** **(c)airplane model**

Figure 7. The test colored model

The thresholds and benchmark vectors used in the experiment are computed based on the colored dataset in Figure 3. During computing the threshold, v is 6, which means that 6 benchmark vectors are computed. The threshold values of the plier and airplane models are 11.1214 and 9.8121 respectively in the experiment. The experimental results are given in Table 1. It can be seen that the

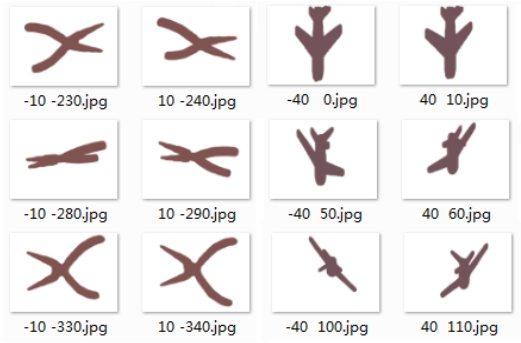


Figure 9. Colored training dataset.

model *b* and *c* are initially divided into the plier's category. Then compare the mean Euclidean distance with the corresponding threshold, which shows that the value of mean Euclidean distance of model *b* is less than the threshold value 11.21 and the mean value of model *c* is greater than the threshold value 11.21. The same result occurs in the model *a*. Therefore, the model *a* and *c* are not trained model and model *b* is trained model which only occurs deformation.

Table 1. Results of model classification.

model	P	W	Q	final result
model *a*	Airplane	18.30	9.81	exclude
model *b*	Plier	6.93	11.21	Plier
model *c*	Plier	16.16	11.21	exclude

Where "P" represents the preliminary classification, "W" represents the mean Euclidean distance, and "Q" represents the threshold value.

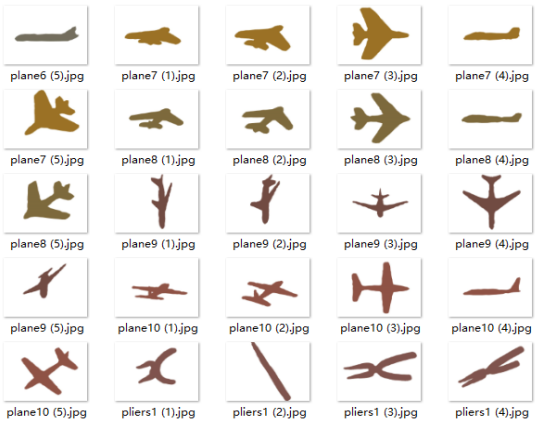


Figure 10. Colored test dataset.

In consequence, the classification used the t-CNN is more accurate that the traditional CNN. Furthermore, the verification in this section only chooses one image as input. Although the result is true, but it is dangerous. Therefore, in section 4, the paper will verify some models in different model datasets and compare our result with the others.

4. EXPERIMENTS

The experiments in the paper are accomplished under Windows 10 64-bit systems, Intel (R) Core (TM) i7-7700 with memory 8GB, GPU for NVIDIA GeForce 1070Ti. In this section, we will compute the result of the recall and precision based on different model sets first. And then compare our retrieval accuracy with the current results.

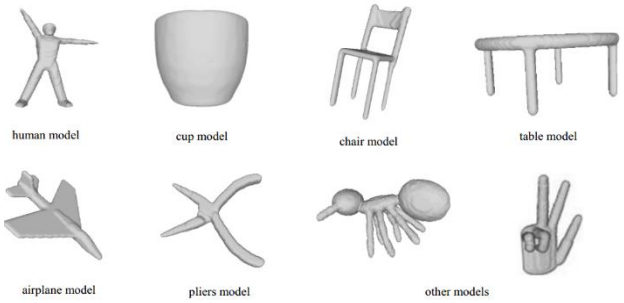


Figure 8. Different models of McGill model set. The number of models of the set is 505. Meanwhile, there are deformed models in each category.

The paper chooses the whole models in the McGill model set [10] as the experimental object and compare the accuracy of the results in [11], which the number of models is larger than the number of models used in [11]. Part of models of the McGill model set is shown in Figure 8.

Construct a colored training dataset and a test dataset used the models in Figure 8 based on the viewpoint trackball algorithm. The recall and precision [12] will be used to verify the accuracy of the classification result. To verify the effectiveness of the t-CNN, especially the threshold, 4 airplane models are selected from total 26 airplane models and one human model is selected as the experimental model. Constructed colored training set is shown in

Figure 9. A colored test dataset is constructed simultaneously used the models of the entire McGill model set (in Figure 10).

Firstly, a five-category t-CNN was trained used the colored training dataset. Through the trained t-CNN, the threshold values of five categories of models are computed, which the four airplanes are 40.0678, 19.9535, 7.1175 and 10.69945 respectively according to the mark number of the model. The threshold value of the human model is 12.27905. The numbers v of images for computing the threshold are 6. Images of test dataset in Figure 10 are put into the trained t-CNN, where each model selects 5 images as input and a mean value of Euclidean distance **W** can be obtained. Parts of experimental results are shown in table 2.

Secondly, the airplane models 1, 3, 6, and 9 have the correct categories after comparing with the corresponding thresholds, which can see the underlined values of the Table 2. The images of the airplane model 5, which are not trained in the t-CNN, are divided into the two categories first. The two images are belong to the second category and the other three images are divided into the fifth category. According to the steps in the verification stage, the two average values of Euclidean distance are 28.2060 and 18.4913, which they are greater than the corresponding category threshold so that this model should be excluded from the two categories. However, all images of Human model 8 are initially classified into the second category. And the mean value 17.1514(mark in red) of Euclidean distance is less than the threshold value 19.9535 of the second category. Thus, the classification is error.

Thirdly, the paper compares the retrieval result with [11] and returns the 10 most similar airplane models. Due to the t-CNN only use four categories of airplane models as classification models, some not trained models will be regarded as the most similar category. Furthermore, models of the entire McGill model set are used for testing the t-CNN. Nonetheless, our accuracy of recall and precision is better than the result in [11]. In Figure 11, nine models are retrieved, which the mean value of Euclidean distance is less than the threshold value of the corresponding category. They are arranged according to the mean value of Euclidean distance from the minimum to the maximum. The other models are excluded because of the threshold.

Table 2. Classification results of the five models.

A / B	**Plane1**	**Plane9**	**Plane3**	**Plane6**	**Huma1**
Plane1	P <u>36.0695</u>	-	-	-	-
Plane3	-	-	P <u>5.5433</u>	-	-
Plane5	-	P (2 imgs) 28.2060	-	-	P (3 imgs) 18.4913
Plane6	-	-	-	P <u>10.4532</u>	-
Plane9	-	P <u>10.9825</u>	-	-	-
Plane10	-	-	-	P 24.0451	-
Birds4	-	P 44.0945	-	-	-
Human 8	-	P 17.1514	-	-	-

Where "A" represents the training model in row 1, "B" represents the test model in column 1 and "P" represents the preliminary classification.

Figure 11. Retrieval results of the airplane

Figure 12 shows the retrieval results in [11]. It can be seen from the above experiment that the model retrieved by the method in [11] not only has an airplane model, but the last row is the bird model. The retrieval results in our paper are airplane models except the images 6. In order to compare with the experimental results in [11], the calculations of recall and precision are computed according to the 10 models. Combined with the threshold, 9 models are retrieved in our result. The percentage of precision is 88.89%, and the recall rate is 80%, while the percentages of precision and recall are only 40% in [11]. It can be seen from the experiment that the experimental results of the paper are more accurate. Even if the percentage of precision is only 88.89%, a better result will be achieved when a bigger t-CNN for more category classification is trained.

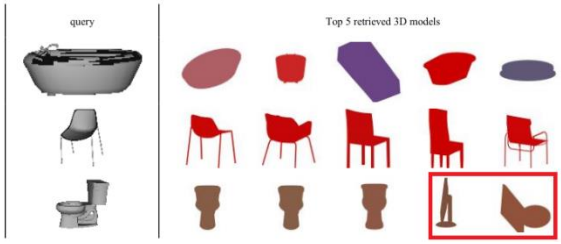


Figure 13. 3D models of ModelNet10 retrieval examples. Top matches are shown for each query, with mistakes highlighted in red.

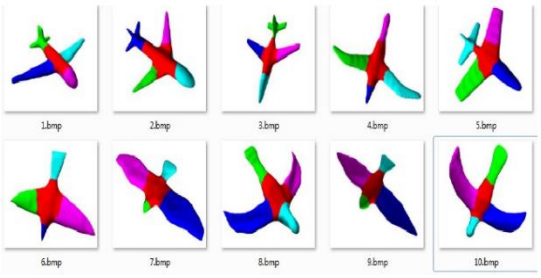


Figure 12. Retrieval results of the airplane models in [11].

Finally, the ModelNet10 [13] are tested in our paper, which it is frequently used model set in current papers. A t-CNN with 10 categories classification is trained used the colored dataset. Select one model of each category as the training model. Whole models in ModelNet10 are the test models. And the threshold value of each category is computed by selected 6 images. Similarly, during the retrieval, 6 images of each test models are generated by the viewpoint trackball algorithm. Then the top 5 retrieved 3D models of each category are returned, as shown in the Figure 13.

Since the majority of models have the right classification result, there are only shown the three models, which the error result occurs in the retrieval of toilet model. In Figure 13, one model of each category of the ModelNet10 is chosen for training the t-CNN. The retrieval result with a threshold is shown that only the training models are the true model, for instance, the mean value of Euclidean distance of the first model bathtub model is 0.640641, which is less than the threshold value of 1.737031. However, the value of the second retrieval bathtub model is 27.71252, which is much larger than the threshold value of the trained bathtub model. Therefore, if we only want to find the trained model in the model set, the retrieval result is also exactly. However, we should compare our result with the other methods. Hence, other retrieval models are arranged according to the mean value of Euclidean distance from the minim to the maximum. The retrieval results of different methods are displayed in the table below. As can be observed in the table 3, our method is excellent. In [14], the PANORAMA-ENN method can obtain a higher accuracy than our method, which uses the 2D panoramic view representation of 3D models as input to an ensemble of CNN. The 2D image processing is also considered in the paper [14]. However, the more images should be put into the CNN in PANORAMA-ENN method. This means that if we increase the input images, we also can get a more excellent result. Because of, if we do it, the more shape features will be learned in the t-CNN.

Table 3. Classification results on ModelNet10.

Method	ModelNet10 Classification (Accuracy)
3DShapeNets [13]	83.50%
Geometry Image [15]	88.40%
LightNet [16]	93.94%
MLH-MV [17]	94.80%
t-CNN(ours)	96%
PANORAMA-ENN [14]	96.85%

5. CONCLUSION

The similarity analysis, retrieval and matching of 3D models have always been an important research point, but there is still no convenient and accurate 3D model retrieval method. Even with the advent of deep learning technology, there is still no highly accurate 3D model retrieval algorithm. The problems of the 3D models have not been solved. In our paper, we consider how it is easy to put 3D model data into a CNN while preserving the global and local features of the model as much as possible. At the same time, for model retrieval, in order to guarantee the accuracy of the retrieval, the structure of CNN is updated, and the threshold is proposed. On this basis, this paper proposes the t-CNN method for 3D model similarity analysis. It can be seen from the experiments that the method is excellent. However, there are some uncertain factors in our method. For example, the threshold value is computed by *v* images of the 3D models. The number *v* of images may be affecting the threshold value. If we increase the number of images, the threshold will decrease, which means that the retrieval is more demanding and the final result will also be affected. Similarly, the number *l* of images in the test stage will affect the final category confirmation. Therefore, how to compute the threshold and the number of images selected will be the focus of our future research work.

6. ACKNOWLEDGMENTS

This research was supported by National Natural Science Foundation of China under Grant No.11671009, the Department of Education of Guangdong Province under Grant No.2017KQNCX275, and the South China Institute of Software Engineering of Guangzhou University under Grant No.Ky201726.

7. REFERENCES

[1] Xie, Z., Xu, K., Liu, L., and Xiong, Y. 2015. *3D shape segmentation and labeling via extreme learning machine.* Computer Graphics Forum, 33(5), 85-95. DOI= https://doi.org/10.1111/cgf.12434.

[2] Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. 2015. *Multi-view convolutional neural networks for 3D shape recognition.* ICCV. DOI=https://doi.org/10.1109/ICCV.2015.114.

[3] Wang, P. S., Liu, Y., Guo, Y. X., Sun, C. Y., and Tong, X. 2017. *O-cnn: octree-based convolutional neural networks for 3D shape analysis*. Acm Transactions on Graphics, 36(4), 72.DOI = https://doi.org/10.1145/3072959.3073608.

[4] Charles, R. Q., Su, H., Mo, K., and Guibas, L. J. 2016. *Pointnet: deep learning on point sets for 3D classification and segmentation.* 77-85. DOI = http://doi.ieeecomputersociety.org/10.1109/CVPR.2017.16.

[5] Masci, J., Boscaini, D., Bronstein, M. M., and Vandergheynst, P. 2015. *Geodesic convolutional neural networks on riemannian manifolds.* 832-840.

[6] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. *ImageNet Classification with Deep Convolutional Neural Networks*. International Conference on Neural Information Processing Systems. DOI= https://doi.org/10.1145/3065386.

[7] Smarajit, B., Amita, P., Disha C., and Taranga M. 2017. *Improved Content-Based Image Retrieval via Discriminant Analysis*. International Journal of Machine Learning and Computing. 7(3), 44-48. DOI= 10.18178/ijmlc.2017.7.3.618.

[8] Brock, A., Lim, T., Ritchie, J. M., and Weston, N. 2016. *Generative and discriminative voxel modeling with convolutional neural networks.* Computer Science.

[9] Luo, L., Ruan, W.,Zhuang, J.,Liu, W., and QIN, S. 2018. *Similarity Analysis of 3D Rigid and Non-rigid Models Based on CNNs.* Software Guide.

[10] Siddiqi, K., Zhang, J., Macrini, D., Shokoufandeh, A., Bouix, S., and Dickinson, S. 2008. *Retrieving articulated 3-D models using medial surfaces.* Machine Vision and Applications, 19(4), 261-275. DOI = https://doi.org/10.1007/s00138-007-0097-8.

[11] Liu,F.S. 2016. *Research on 3D Mesh Clustering Segmentation and Retrieval Method Based on Salient Points.* (Master dissertation) Jilin University.

[12] Wikipedia contributors. Precision and recall [EB/OL].[2018].https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=853202943.

[13] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and J. Xiao. 2014. *3D shapenets: a deep representation for volumetric shapes*. DOI = http://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298801.

[14] Sfikas, K., Pratikakis, I., and Theoharis, T. 2017. *Ensemble of panorama-based convolutional neural networks for 3D model classification and retrieval.* Computers & Graphics.

[15] Sinha, A., Bai, J., and Ramani, K. 2016. *Deep Learning 3D Shape Surfaces Using Geometry Images.* Computer Vision – ECCV 2016. Springer International Publishing. DOI = https://doi.org/10.1007/978-3-319-46466-4_14.

[16] Xu, X., and Todorovic, S. 2017. *Beam search for learning a deep Convolutional Neural Network of 3D shapes.* International Conference on Pattern Recognition .3506-3511. IEEE.

[17] Sarkar, K., Hampiholi, B., Varanasi, K., and Stricker, D. 2018. *Learning 3D shapes as multi-layered height-maps using 2D convolutional networks.*

A Simple Image Acquisition System and Its Calibration for Image-Based 3D Reconstruction

Xiaoming Wang
Department of Mathematics and Physics, Shandong Jiaotong University
Jinan, China,250300
wxm_sd@126.com

Limin Shi
Institute of Automation Chinese Academy of Science
Beijing,China,100190
limin.shi@ia.ac.cn

Huarong Xu
Department of Computer Science and Technology, Xiamen University of Technology,
Xiamen, china, 361024
hrxu@xmut.edu.cn

ABSTRACT
This paper presents a simple image acquisition system composed of multiple cameras and a turntable, through which object's multiple-view images can be acquired and calibrated automatically. The main contribution of this paper is proposing a calibration algorithm for this system which using a simple calibration equipment to calibrate the cameras' intrinsic and their extrinsic relate to the center of turntable, and with these system parameters, all the acquired multi-view images can be easily calibrated. Combined with MVS, mesh generation and texture mapping algorithms, complete texture model can be automatically reconstructed from the calibrated images. The calibration and reconstruction examples based on the system are given in the end of the paper. The experimental results show the presented system with our calibration algorithm has high reconstruction accuracy, and can be used for many reconstruction applications.

CCS Concepts
• Computing methodologies → Reconstruction • Computing methodologies → Camera calibration.

Keywords
Camera calibration; Image-based 3D reconstruction.

1. INTRODUCTION
3D reconstruction of a scene from multiple stereo views is called image-based modeling (IBM) technique, which is important for diverse applications ranging from robotics vision, electronic earth maps, and virtual reality to 3D film production, computer games and animation.

A complete image-based 3D reconstruction process usually includes four steps: image calibration, MVS, mesh generation and texture mapping. In recent years, with the in-depth research, many practical algorithms continue to be raised, such as image calibration: bundler[1], Photometric BA [2]; MVS: PMVS[3], Patchmatch based MVS[4], Colmap[5]; mesh generation: Poisson reconstruction[6], graph-cut based surface reconstruction[7]; texture mapping: seam levelling[8], Seamless montage[9], Large-Scale texturing[10], etc. Some authors of these works have opened their source codes to the public, and many excellent algorithms have been integrated into practical open source libraries, such as OpengMVG[11] and OpenMVS[12]. All these have greatly promoted the development of IBM. In addition, as technology continues to mature, some commercial software systems are also emerging and have been widely applied in many fields, for example PhotoScan, Capture Reality, Smart3D, and so on.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12 $15.00
https://doi.org/10.1145/3301506.3301543

3D reconstruction of an object from multiple stereo views usually need to take tens or even hundreds of pictures around the object to cover the whole surface of the object and ensure the overlap rate among the images. It is a very tedious process. For 3D reconstruction, after the images captured, images calibration is the first job. For some weakly textured objects, due to the fewer feature points, it is difficult to complete calibrate all the images by self-calibration algorithms which usually lead to failure of reconstruction.

In view of the above problems, a simple data acquisition system based on turntable and camera array is designed and proposed in this paper. The acquisition system has no complex mechanical system, and it can set the shooting orientation of every camera, thus ensuring the coverage of the image to the surface of the object and the overlap rate among images. In addition, a simple and automatic calibration algorithm is also presented based on the characteristics of the system which greatly reduces the difficulty of reconstruction for weak texture objects. The main work of this paper, combine with the existing MVS，mesh generation and texture mapping algorithms, can form a full-automatic image acquisition and reconstruction system.

The remainder of this paper is organized as follows. in Sec. II, the image acquisition system is briefly reviewed. Then the calibration algorithm is presented in Sec. III and the performance of the image acquisition system and our calibration algorithm are demonstrated in Sec. IV. A short conclusion is presented at the end of this paper.

2. SYSTEM INTRODUCTION
Seits et al.[13] captured images using the Stanford Spherical Gantry to generate multi-view dataset bench marks which enables moving a camera on a sphere to specified latitude/longitude angles. In addition, they used Jean-Yves Bouguet's matlab toolbox and their own software to calibrate the cameras. But the Stanford Spherical Gantry is a high elaborate mechanical device, the cost is higher. In addition, their calibration process for this system is complicated.

In order to automatically achieve multi angle images, this paper design a relatively simple image acquisition system which includes a hardware system and a software system. The hardware system includes a multi-camera array and a turntable, as shown in figure 1(a). The camera array is arranged vertically by a number of cameras fixed to a bracket which at a certain distance from the turntable. The software system is a control system which control all cameras exposure one time after turntable rotate a set angle. Suppose there are m cameras in the camera array, and the set rotation angle is ω. after the turntable has rotated one week, the acquired image sequence by camera C_i is $I_i = \{I_i^j | j = 1,2,\cdots,n\}$, where $i = 1,2,\cdots,m, n = \lceil 360/\omega \rceil$. In this paper, m=3 is used which can satisfy 3D reconstruction of a large class of objects.

The image acquisition system is designed for 3D reconstruction of objects. For different shapes or sizes of objects, in order to capture the images as far as possible to cover the entire object, cameras' positions, orientation, focal lengths can be adjusted at will. The basic adjustment principle is that the camera located on the upper end of bracket tilts downward to capture the images of object top; the camera located in middle remains relatively horizontal, mainly captures the images of object side, and the camera located on the lower end of bracket tilts upward to capture the images of the object bottom.

After setting up the cameras' orientation and positions, the geometric relationship between every camera and the turntable is relatively fixed. We can realize the automatic calibration of the images through the calibration algorithm presented in the next section.

a b

Figure 1. Our image acquisition system (a) and hand-made calibration equipment used in the experiments(b).

3. ALGORITHMS

The calibration algorithm is a main contribution of this paper. In this section we will give the details of our system calibration algorithm.

3.1 System Calibration

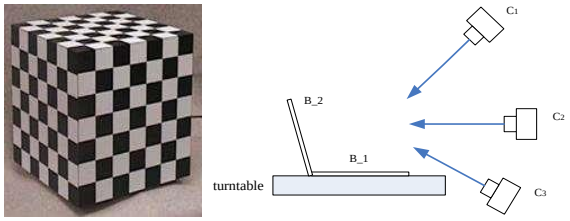
Images calibration is the first step of image-based 3D reconstruction. As our image acquisition system is relatively fixed, we will firstly calibrate the system parameters which include cameras' intrinsic parameters and their extrinsic parameters relative to the rotation center of turntable and then use the system parameters to achieve automatic images calibration.

We set the intrinsic matrix of camera C_i is K_i, and its extrinsic parameters relative to the rotation center of turntable is R_i、T_i, R_i is rotation matrix，T_i is translation vector, $i = 1,2,3$.

Because of the large difference in cameras' orientations, Multi-camera calibration of above acquisition system usually requires special calibration blocks that the accurate geometric location of any corner of the chessboard on every face is known, as shown in figure 2(a). Such calibration block is usually expensive. This paper designed a simple calibration equipment which include two different chessboards, as shown in figure 1(b). Suppose B_1, B_2 are two chessboards. B_1 is horizontally parallel to the turntable plane. The angle between B_1 and B_2 is equal to or slightly greater than 90 degrees, which is not need to be known exactly. that allows the calibration equipment can be made by hand.

Due to cameras' orientations limitation described earlier. B_1 can not be seen by the bottom camera, and its images captured by the intermediate camera will be larger stretched. B_2 can be seen by all cameras. Only the top camera has good orientation for capture image of B_1 and B_2 simultaneously. As shown in figure 2(b). In our calibration algorithm, B_1 is first used to compute the center of turntable, then all cameras' intrinsic and extrinsic parameters are calibrated based on the inherent geometric relationship between B_1 and B_2.

Place the calibration equipment on the turntable and run the turntable, every camera takes one picture simultaneous at every time turntable rotate a certain angle ω. suppose $I_i^1, I_i^2, \cdots, I_i^s$ are the acquired images by camera $C_i, i = 1,2,3$. The following, we'll use these images to calibrate the system parameters.



a b

Figure 2. Calibration blocks usually used for multi-camera calibration(a) and Schematic diagram of our image acquisition system(b).

3.1.1 Compute center of turntable

The center of turntable is a fixed point when the turntable rotating. In order to find this position as precise as possible，the following method is designed.

From multi-view geometry[14], there exists a homography between B_1 and every its image. Detect corners of chessboard B_1 from $I_1^1, I_1^2, \cdots, I_1^s$, suppose $[x_j^i, y_j^i]$ is the jth corner of the chessboard B_1 detected in I_1^i, $[X_j^i, Y_j^i]$ is the world coordinate corresponding to image pixel, where $i = 1,2,\cdots,s$, $j = 1,2,\cdots,n1$, n1 is the number of corners detected in B_1. Then exists a homography $\mathbf{H}$ between $[x_j^i, y_j^i]^T$ and $[X_j^i, Y_j^i]^T$. Initialize $[X_j^1, Y_j^1]$ as $\left[\left(j - \left[\frac{j}{w}\right]w\right)h, \left[\frac{j}{w}\right]h\right]^T$ where $\mathbf{h}$ is the length of squre in chessboard, and $\mathbf{w}$ is the number of squres every row of chessboard in B_1.

Firstly, we compute the homography H from $[x_j^1, y_j^1]^T$ to $[X_j^1, Y_j^1]^T$ for all $j = 1,2,\cdots,n1$.

$$\begin{bmatrix} X_j^1 \\ Y_j^1 \end{bmatrix} = H \begin{bmatrix} x_j^1 \\ y_j^1 \end{bmatrix} \qquad (1)$$

Since the freedom degrees of H is 8, To solve H from (1), only need $n1 \geq 8$. If $n1 > 8$, the RANSAC[14] is used to robust estimate the least square solution of H.

Then based on the following equation(2), $[X_j^i, Y_j^i]^T$ can be computed, $i = 2,3,\cdots, s; j = 1,2,\cdots n1$.

$$\begin{bmatrix} X_j^i \\ Y_j^i \end{bmatrix} = H \begin{bmatrix} x_j^i \\ y_j^i \end{bmatrix} \quad (2)$$

From the constant rotation angle ω of the turntable, we also have

$$\begin{bmatrix} X_j^i - X_0 \\ Y_j^i - Y_0 \end{bmatrix} = R(\omega) \begin{bmatrix} X_j^{i-1} - X_0 \\ Y_j^{i-1} - Y_0 \end{bmatrix}$$

which is

$$(I - R(\omega)) \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} = \begin{bmatrix} X_j^i \\ Y_j^i \end{bmatrix} - R(\omega) \begin{bmatrix} X_j^{i-1} \\ Y_j^{i-1} \end{bmatrix} \quad (3)$$

where $R(\omega)$ is a rotation matrix

$$R(\omega) = \begin{bmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{bmatrix}$$

Construct equation as (3) for every $i = 1,2,\cdots, s$, $j = 1,2,\cdots, n1$, then the least squares solution of rotation center $[X_0, Y_0]$ can be solved.

From the above derivation, s only need larger than 1, but for more accurate, we use $s = 6$ in the experiments.

3.1.2 Camera calibration

Since B_2 is seen in all images, based on the chessboard corner points of B_2 extracted from I_i^j, $i = 1,2,3$, $j = 1,2,\cdots s$, Zhang's plane calibration method [15] can be used to get the camera intrinsic K_i and extrinsic parameters related to B2 $\bar{R}_i$、$\bar{T}_i$, $i = 1,2,3$. Then the extrinsic $R_{i\sim1}$、$T_{i\sim1}$ that camera C_i relative to camera C_1 can be computed as follows, $i = 2,3$:

$$R_{i\sim1} = \bar{R}_i \bar{R}_1^T$$
$$T_{i\sim1} = \bar{T}_i - R_{i\sim1} \bar{T}_1$$

We also can see $(I_1^j, I_i^j), j = 1,2,\cdots s$ as stereo pairs and use stereo calibration to get $R_{i\sim1}$、$T_{i\sim1}$ respectively for all $i = 1,2,3$.

Zhang's plane calibration algorithm [15] and stereo calibration are all very popular, this paper does not describe them in detail.

Using the rotation center $[X_0, Y_0]$ computed as above, the origin of world system can be transformed to $[X_0, Y_0]$ by

$$\begin{bmatrix} X'^j_i \\ Y'^j_i \end{bmatrix} = \begin{bmatrix} X_i^j \\ Y_i^j \end{bmatrix} - \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}$$

Then the following corresponding between 3D Points $[X'^j_i, Y'^j_i, 0]^T$ and 2D image points $[x_j^i, y_j^i]^T$ is established for all $i = 1,2,3, j = 1,2,\cdots n$.

$$\lambda \begin{bmatrix} x_i^j \\ y_i^j \\ 1 \end{bmatrix} = K_1 [R_1 | t_1] \begin{bmatrix} X'^j_i \\ Y'^j_i \\ 0 \end{bmatrix}$$

Using PnP method[16], we can easily get the extrinsic parameters R_1, T_1.

Then R_i, T_i $i = 2,\cdots, m$ can be computed as follows

$$R_i = R_{i\sim1} R_1$$
$$T_i = R_{i\sim1} T_1 + T_{i\sim1}$$

At this point, the camera's intrinsic parameters and the extrinsic parameters relative to the center of the turntable are all calibrated, and system calibration is finished.

3.2 Images Acquisition and Calibration

Place the object on the rotation platform, and all the camera acquire an image simultaneous at every time turntable turn a certain angle ω. suppose $I_1, I_2 ..., I_n$ are the acquired images after the platform rotated a week.

Notice the turntable rotating can be considered the turntable is fixed and camera array rotates around the center of rotation platform in the opposite direction. Suppose R_i^j 、 t_i^j are the extrinsic parameters of image I_i^j, using the system parameters get from the former section, R_i^j、t_i^j can be computed by the following equations

$$R_i^j = R_i * R^T(\theta_j)$$
$$t_i^j = t_i$$

Where $R(\theta_j)$ is a rotation matrix around Z-axis of world system by an angle $\theta_j, \theta_j = j * \omega, j = 1,2,\cdots n$.

Combined with the camera's intrinsic parameters, we can get the following projection matrix of I_i^j for all $i = 1,2,3, j = 1,2,\cdots n$,

$$P_i^j = K_i [R_i * R^T(\theta^j) | t_i]$$

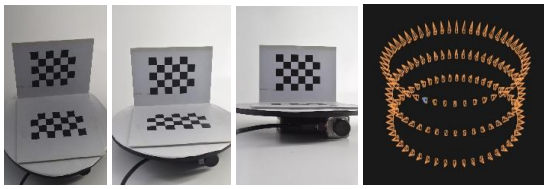
3.3 3D Reconstruction Based on Calibrated Images

Based on the acquired calibrated images, we can use the existing MVS, mesh generation and texture mapping algorithms to reconstruct the 3D model of object. Among many MVS algorithms, PMVS[3] and Colmap[5] are two of best performing algorithms. In this paper we use PMVS to reconstruct dense oriented point cloud. Mesh surface reconstruction from point cloud is a traditional research direction in CG. A lot of algorithms have been proposed [6][7][16]. In these methods, Poisson reconstruction[6] is one of the best known algorithms due to its good performance in many applications. So this paper uses Poisson reconstruction to get mesh model from oriented dense point cloud. Finally, texture mapping algorithm[10] is use to create model texture. In the next section, we will combine these algorithms to test the practicability of our acquisition system and calibration algorithm presented above.

4. EXPERIMENTS

Since the particularity of the system and calibration algorithm, there is no existing corresponding algorithm and data that can be used to compare with our algorithm. This paper only illustrates the practicability of the system and the effectiveness of the algorithm through several sets of real data experiments.

We firstly test the proposed calibration algorithm on real data. In our experiments $\omega = 12^o$, m=3, s=6 and $n1 = 6 * 3 = 18$. Some captured images of calibration equipment in the experiments are shown in figure 3(a).The calibration algorithms are developed by OpenCV. Mean reprojection errors are shown in Table 1. From Table 1, we can see our calibration algorithm has high accuracy.



a b

Figure 3. Image of our calibration equipment captured by C1、C2 and C3 respectively (a). Images calibration results by our calibration algorithom(b).

Table 1. accuracy of calibration and reconstruction

	Mean error	Max error	Min error
Reconstruction error (mm)	0.286467	0.527317	0.190325
Reprojection error (pix)	0.437765	1.074330	0.272254

Some real objects are used to test the practicability of our acquisition system and calibration algorithm for image-based 3D reconstruction. Firstly, multi-view images of object are taken automatically by our image acquisition system. Then the calibrated system parameters are used to calibrate every image. Finally, texture model is generated by the procedure presented in Section 3.3. Three of the reconstructed results are shown in figure4, 5. Mean reconstruction errors are also shown in Table 1. Since the images cover almost all parts of the object, the reconstructed model is nearly complete. In addition, based on system parameters calibrated in advance, all images can be calibrated automatically, as shown in Figure3(b). Some weak texture objects that can't be reconstructed by self-calibration methods like bundler can be reconstructed well by our system, such as Mayun's cartoon statue shown in figure 5.

a b c

Figure 4. The three-camera-array image acquisition system in our experiments(a), some multi-view images acquired by the cameras(b) and the final reconstructed texture model(c)

Figure 5. Some reconstruction results in our experiments. Two of the acquisited images(a)、(b);The reconstructed models(c).

5. CONCLUSION

This paper presents an image acquisition system and corresponding calibration algorithm for objects reconstruction using camera array and a rotation platform. The main advantages of our system and algorithm include: (1) automatic images acquisition;(2) capturing the images in as many views as possible to ensure the complete of the reconstructed model; (3) automatic images calibration; (4) no need to extract feature points to calibrate the image, so some weak texture objects can be reconstructed well.

The data acquisition system and calibration algorithm proposed in this paper can be combined with the existing MVS, mesh generation and texture mapping algorithms to form a full-automatic acquisition and reconstruction system.The experiments presented in the end of paper demonstrate our algorithm has high calibration accuracy, and the reconstructed results based on the calibrated images by our system have both high completeness and precision.

6. ACKNOWLEDGMENTS

This work was supported by National Nature Science Foundation of China under the research project 61772444, and Fujian province Science and Technology Department under the research project 2018I0026.

7. REFERENCES

[1] Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3D. In: SIGGRAPH (2006)

[2] Delaunoy, Amaël and Pollefeys, Marc.: Photometric Bundle Adjustment for Dense Multi-view 3D Modeling , In:CVPR (2014).

[3] Yasutaka Furukawa and Jean Ponce Accurate, Dense, and Robust Multi-View Stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, Issue 8, Pages 1362-1376, August 2010.

[4] Zheng, E., Dunn, E., Jojic, V., Frahm, J.M.: Patchmatch based joint view selection and depthmap estimation. In: CVPR (2014).

[5] Schönberger, Johannes L. and Zheng, Enliang and Frahm, Jan-Michael and Pollefeys, Marc.:Pixelwise view selection for unstructured multi-view stereo, ECCV (2016).

[6] M. Kazhdan, M. Bolitho, and H. Hoppe, Poisson surface reconstruction. Proceedings of the fourth Eurographics symposium on Geometry processing, 2006.

[7] J. Pons, R. Keriven and P. Labatut.:Efficient Multi-View Reconstruction of Large-Scale Scenes using Interest Points, Delaunay Triangulation and Graph Cuts. ICCV 2007.

[8] Lempitsky, V., Ivanov, D.: Seamless mosaicing of image-based texture maps. In: CVPR (2007).

[9] Gal, R., Wexler, Y., Ofek, E., Hoppe, H., Cohen-Or, D.: Seamless montage for texturing models. Computer Graphics Forum 29 (2010).

[10] Michael Waechter,Nils Moehrle,Michael Goesele.:Let There Be Color! Large-Scale Texturing of 3D Reconstruction. ECCV 2014.

[11] https://openmvg.readthedocs.io/en/latest/

[12] https://github.com/cdcseacave/openMVS/wiki/Building/

[13] Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison andevaluation of multi-view stereo reconstruction algorithms.In: CVPR(2006).

[14] Hartley R, Zisserman A. Multiple View Geometry in Computer Vision (Second edition). London: Cambridge University Press, 2004.

[15] Z. Zhang. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330-1334, 2000.

[16] ZHAO H.-K., OSHER S., FEDKIW R.: Fast Surface Reconstruction Using the Level Set Method. In First IEEE Workshop on Variational and Level Set Methods (2001), pp. 194-202.

Chapter 4: Image Analysis

Omnidirectional Saliency Map Generation by Yin-Yang Grid Method

Daiki Okazaki, An-shui Yu, Kenji Hara
Department of Visual Communication Design
Kyushu University
Shiobaru 4–9–1, Minami-ku, Fukuoka, 815-8540 JAPAN
kysu201210@gmail.com

ABSTRACT
In this paper, we extend the existing saliency map generation methods to deal with fisheye images. Our method makes use of an overset grid comprising two latitude-longitude grids. We experiment with real fisheye images and demonstrate the effectiveness of our proposed method.

CCS Concepts
•Computing methodologies → Image processing • Theory of computation → Convex optimization.

Keywords
saliency map, fisheye image, overset grid

1. INTRODUCTION
A saliency map is an evaluation of the saliency for each pixel to extract a region or an object of interest from a still image or a video image. Saliency map generation can be applied to various fields such as gaze analysis, object detection, robot vision, among others and research has been actively carried out in this regard [1]-[6]. In this paper, we deal with the problem of estimating the saliency map of omnidirectional images. Various methods of saliency map generation have been proposed in the literature. However, most of them are tailored for images with normal viewing angles and uniform spatial resolution. Consequently, a wide range of luminance information around 360° is dismissed as not appropriate to use it. This is due to the fact that wide viewing angle images present uneven spatial resolution. This is the case of omnidirectional images. There is also the so-called pole problem. Normal image processing calculation becomes impossible or unstable in the vicinity of poles or singular points of an omnidirectional image. Furthermore, the viewing angle of the human eyes is only about 200°. It is thus difficult to collect learning data by gaze measurement using 360° omnidirectional images. For this reason, learning approaches such as deep learning are difficult when targeting omnidirectional images.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301508

In this paper, we extend the existing saliency map generation methods for ordinary narrow view image so that it can be applied to omnidirectional images. In the proposed method, to cope with resolution nonuniformity and the pole problem in omnidirectional images, we use an overset grid comprising two latitude and longitude grids [7]. The geometric distortion is small and the resolution is spatially uniform from the omnidirectional image. Two rectangular planar images can thus be generated. The advantage of this method is that existing code of arbitrary saliency map generation method can be reused without modification. The proposed method is applied to actual omnidirectional images and its effectiveness is demonstrated.

2. RELATED METHOD
Many saliency map generation methods have been reported. For example, there are numerous methods starting from a computational model of saliency map based on characteristic theory [2]; a model [3] based on a learning-based approach; and others [4]-[6]. However, if these methods are applied directly to omnidirectional images, it is not possible to generate accurate saliency maps due to nonuniformity in resolution and polar problems of omnidirectional images. For these methods, Bogdanova et al.'s salience map generation method [8] is a natural extension of Itti et al.'s method [2] for omnidirectional images. It can generate an accurate saliency map for processing on a spherical surface without geometric distortion. However, their method is computationally complicated due to the required processing on the sphere. The estimation accuracy remains unchanged with respect to that of Itti et al.'s method. The proposed method has the advantage that it can easily generate the saliency map by reusing the previous method while responding to the resolution unevenness and the polar problem of omnidirectional images.

3. PROPOSED METHOD
We next describe the details of our method for generation of the saliency map from all the circumference luminance data obtained by photographing the surrounding 360° omnidirection. Fig. 1 shows the procedure of the proposed method. The rectangular image of $\theta \in [0,\pi]$, $\phi \in [0,2\pi)$ obtained by coordinate transformation to the spherical coordinate system is depicted in Fig. 1(a). However, the image size is 1024 × 1024.

3.1 Yin Yang Grid
In the proposed method, the Yin-Yang grid of Kageyama et al. [7] is used. In recent years, Yin Yang grid was devised as one of the solutions to the polar problem arising from the fact that Earth is spherical within the Earth science field. It is a Yin-grid consisting of the low latitude region of latitude and longitude lattices, and a Yang-grid in which the Yin-grid was rotated so as to cover the

high latitude region. The overset grid on the spherical surface partially overlaps the spherical surface (Fig. 2). By using this Yin-Yang grid for omnidirectional panoramic image processing, as in the field of Earth science, we preclude polar problems. Fig. 1 (b) shows a Yin-grid image corresponding to the Fig. 1 (a) (upper part of Fig. 1 (b)) and Yang-grid image (bottom part of Fig. 1 (b)). As the geometric distortion of these two rectangular images has been corrected, existing saliency map generation techniques can be applied unchanged (Fig. 1 (c)).

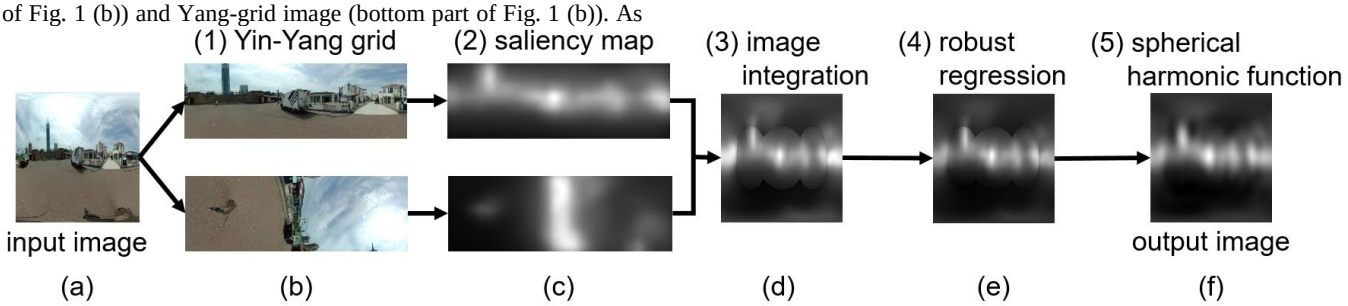


Figure 1. Information flow of our method.

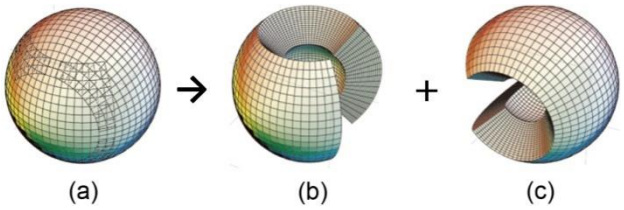


Figure 2. Yin-Yang grid.

(a) Yin-Yang grid, (b) Yin-grid, (c) Yang-grid. ([7]).

3.2 Saliency Correction Based on Overlapping Region

Image stitching is performed based on the overlapping area of the saliency map of both the Yin-grid image and the Yang-grid image. First, the saliency map of each grid image is integrated from the coordinate system of each grid back to the representation of the coordinate system of the panoramic image. At this time, the pixel value of the saliency map of the Yang-grid image is used in the overlapping area. The result after the integration is depicted in Fig. 1 (d). Next, in order to match the saliency level of the Yang-grid image to the Yin-grid image, robust regression is performed based on the pixel value of the corresponding overlapping region when matching the coordinate system of the Yin-grid image. In the proposed method, we formulate an optimization model that estimates the tone mapping function in the L1 norm as follows.

$$\min_{p_1, p_2} ||\mathrm{x}_1 - p_1\mathrm{x}_2 - p_2\mathbf{1}||_1 \quad s.t.\, p_1 > 0, 0 \le p_2 \le T \qquad (1)$$

Here, $\mathrm{x}_1 \in \mathbb{R}^N$ is a vector of pixel values of the overlapping region of the Yin-grid image, $\mathrm{x}_2 \in \mathbb{R}^N$ is a vector of pixel values of the overlapping region of the Yang-grid image, $\mathbf{1}$ is an N-dimensional vector in which all elements are 1, N is the number of pixels in the overlap region, p_1 is a linear coefficient of the tone mapping function, and p_2 is an intercept of the tone mapping function. The vector of the pixel values of each overlap region is normalized by [0, 1]. The reason why we constrain the linear coefficient p_1 to positive values is that the tone mapping function must be a monotonically increasing function. Furthermore, in order to prevent the intercept p_2 from becoming a negative value or extreme value, we enforce a restriction such that it must take values larger than 0 and smaller than a certain threshold T. In the proposed method, setting the threshold T to 0.1 prevents the intercept p_2 from becoming extremely large. Since Eq. (1) is a constrained minimization problem, it is difficult to solve just as it is. Therefore, an instruction function is introduced. It is apparently replaced with a minimization problem without

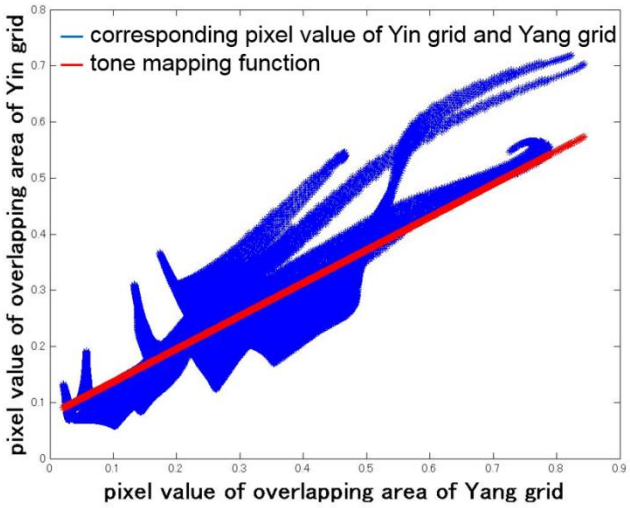


Figure 3. Tone mapping function.

constraints. By doing so, Eq. (1) can be solved using the so-called ADMM (Alternating Direction Multiplier Method). By optimizing Eq.(1), it is possible to obtain the parameters p_1, p_2 where the difference between the pixel value x_2 of the overlapping region of the Yang-grid image and the pixel value x_1 of the overlapping region of the Yin-grid image is minimized. Besides, unlike the L2 norm, excessive dependence on outliers can be avoided by using the L1 norm. The tone mapping function corresponding to p_1 and p_2 obtained in this manner is depicted in Fig.3. Based on this tone mapping function in Fig. 3, after converting the luminance value of the Yang-grid image, the image is normalized (Fig. 1 (e)). However, since unnatural edges still exist at the boundary between the Yin-grid image and the Yang grid image, we solve it by using the spherical harmonic function described in the next section.

3.3 Spherical Harmonic Function

Approximation is performed by using a spherical harmonic function to eliminate discordant elements in the edge portion. The spherical harmonic function can approximately represent an arbitrary function on a spherical surface in the spherical orthogonal basis function system. We must find the optimum scaling factor for each spherical harmonic function to be used. The scaling factor is given by the following equation.

$$c_m^l = \int_0^{\pi} \int_0^{2\pi} f(\theta, \Phi)\, Y_m^l(\theta, \Phi) \sin\theta d\theta d\Phi \qquad (2)$$

Here, l and m are orders of the spherical harmonic function, and vary according to the number of spherical harmonic functions to be used. c_m^l is the factor corresponding to order (l, m), $f(\theta, \Phi)$ is the pixel value at the spherical coordinates (l, m) of the input image. Y_m^l is a function value in the spherical coordinate (θ, Φ) of the spherical harmonic function corresponding to the order (l, m). Approximation is performed by taking the linear sum of the scaling coefficient c obtained by Eq. (2) and the corresponding spherical harmonic function expressed by the following equation:

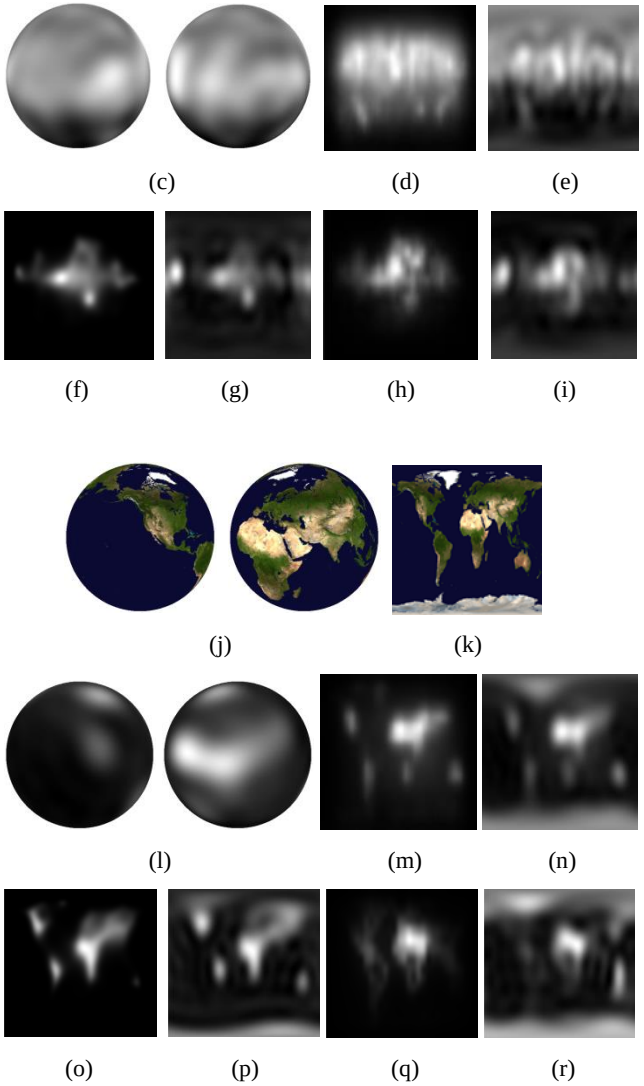
$$S(\theta, \Phi) = \sum_{l=0}^{L} \sum_{m=-l}^{m=l} c_m^l\, Y_m^l(\theta, \Phi) \qquad (3)$$

Here, L is an integer representing the upper limit of the range of degree l, $S(\theta, \Phi)$ is the pixel value of the output image in the spherical coordinate (θ, Φ), that is, S is the saliency map of the omnidirectional image. Accurate approximation can be achieved by setting the maximum order L appropriately. The experiment described in the next chapter, we set $L = 12$. The result of approximating Fig. 1 (e) with a spherical harmonic function is shown in Fig. 1 (f).

4. EXPERIMENTAL RESULTS

We show the result of applying the proposed method to omnidirectional images. The spherical display (Fig. 4(a), (j)) and the panoramic display (Fig. 4(b), 4(k)) indicate the two types of directions in omnidirectional images for input and the 180° direction in the latitude direction, respectively. Although the original size of the panorama image is 1024 × 2048 pixels in total, it is depicted here at 1024 × 1024. We use models of Harel et al. [3], Fang et al. [4], and Tavakoli et al. [6] for generating saliency maps on each grids: (1) Harel's method applied to the panoramic display of each input image (hereinafter referred to as Harel); (2) the method of Fang et al. (hereinafter referred to as Fang); and (3) the method of Tavakoli et al. (hereinafter referred to as Tavakoli) as a conventional method. The high-latitude/low-latitude images obtained by grid division of the proposed method were applied to (4) Harel et al. (hereinafter referred to as "ours-Harel") ; (5) Fang et al. (hereinafter referred to as "ours-Fang") ; to the method of (6) Tavakoli et al. (hereinafter referred to as "ours-Tavakoli"). Performance comparisons were carried out.

(a) (b)



(c) (d) (e)

(f) (g) (h) (i)

(j) (k)

(l) (m) (n)

(o) (p) (q) (r)

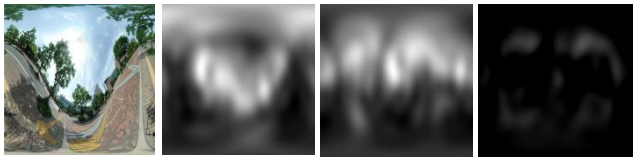
Figure 4. Saliency map generation.

(a) spherical representation of the original image, (b) panoramic representation of the original image, (c) ours-Harel (spherical representation), (d) Harel, (e) ours-Harel (panoramic representation), (f) Fang, (g) ours-Fang, (h) Tavakoli, (i) ours-Tavakoli, (j) spherical representation of the original image, (k) panoramic representation of the original image, (l) ours-Harel (spherical representation), (m) Harel, (n) ours-Harel (panoramic representation), (o) Fang, (p) ours-Fang, (q) Tavakoli, (r) ours-Tavakoli.

4.1 Qualitative Comparison

The saliency maps obtained by the conventional method are shown in Fig. 4 (d), (f), (h), (m), (o), (q), respectively. Fig.4 (e), (g), (i), (n), (p), (r) depicts the panorama of saliency maps obtained by applying the proposed method. Fig. 4 (c) and (l) show the spherical representation of the saliency map obtained by applying the proposed method using Harel et al. In the case of the conventional method, it can be seen that the significance level of the high latitude object region is evaluated to be low. In the image of the world map shown in Fig. 4 (m), (o), (q) , saliency in the area near the polar pole, near the South Pole and the Arctic is not accurately evaluated due to the polar problem. On the contrary, in

the proposed method, the saliency of the object region at high latitude, as well as other regions, can be accurately evaluated.



(a) (b) (c) (d)

Figure 5. Results on coordinate invariance.

(a) input image after the coordinate transformation. Results of saliency map generation：(b) ours-Harel, (c) returning to the original coordinate position, (d) difference image between Fig. 4(e) and Fig. 5 (c).

4.2 Accuracy Evaluation

The results of accuracy evaluation for the proposed method are described next. Fig. 4 (b) is used as the input image. Ours-Harel is used as a saliency map generation method. The results of evaluating coordinate invariance are shown in Fig. 5. The proposed method is applied at ordinary coordinate positions to generate a saliency map (Fig. 4(e)). For the sake of comparison, coordinate transformation is performed so that the center of the input image becomes the position of the overlapping region of the Yin-Yang grid (Fig. 5(a)). The proposed method is applied at the coordinate position to generate a saliency map (Fig. 5(b)). The transformation of the obtained result to the original position is depicted in Fig. 5(c). By comparing each saliency map, coordinate invariance of the proposed method is evaluated. An image obtained by taking the difference between the images of Fig. 4(e) and Fig. 5(c) is shown in Fig. 5(d). From these results, it is understood that the error is very small, and saliency can be estimated with high accuracy regardless of how the coordinate system is considered.

Table 1. Quantitative comparison

Image	Method	RMSE	PSNR [dB]	Correlation coefficient	AUC
Fig.1 (a)	Harel	10.69	16.04	0.63	0.70
	ours-Harel	10.07	22.69	0.94	0.94
Fig.4 (b)	Harel	9.94	14.05	0.47	0.61
	ours-Harel	8.27	24.38	0.88	0.66
Fig.4 (k)	Harel	8.19	24.28	0.87	0.69
	ours-Harel	3.16	34.68	0.97	0.88

The results of the quantitative evaluation are shown in Table 1. The images depicted in Figs. 1 (a), 4 (b) and (k) are used as input images. RMSE and PSNR are used for the evaluation index, AUC is used for the correlation coefficient. For each input image, a difference image is generated by the conventional method and the proposed method, as shown in Fig. 4(d). Harel is used for the conventional method. Here, RMSE and PSNR are evaluated using difference images, and the correlation coefficient uses two saliency maps before the difference. AUC is evaluated by using the lower area of the ROC curve representing the error rate and the detection rate of the difference image on the horizontal axis and the vertical axis respectively. In Table 1, it can be seen that the proposed method is dominant in all evaluation indices, and the robustness to coordinate transformation is high.

5. CONCLUSION

We propose a method to extend the existing saliency map generation method so that it can be applied to omnidirectional images by using an overset grid. Experimental results showed that both the problem of resolution nonuniformity and the polar problem were avoided. Furthermore, we proved and the effectiveness of the proposed method. Future tasks include improving the optimization model to make the robust regression more accurate.

6. REFERENCES

[1] M.M. Chang, G.X. Zhang, N.J. Mitra, X. Huang, and S.M. Hu, Global Contrast *Based Salient Region Detection*, Proc. of CVPR, pp.409–416, 2011.

[2] L. Itti, C. Koch, and E. Niebur, *A model of Saliency Based Visual Attention for Rapid Scene Analysis*, IEEE Trans. PAMI, vol.20, no.11, pp.1254.1259, 1998.

[3] J. Harel, C. Koch, and P. Perona, *Graph-based Visual Saliency*, NIPS, vol.19, pp.545–552, 2006.

[4] S. Fang, J. Li, and Y. Tian, *Learning Discriminative Subspaces on Random Contrasts for Image Saliency Analysis*, IEEE Trans. Neural Netw. Learning Syst.,vol.28, no.5, pp.1095–1108, 2016.

[5] E. Erden, and A. Erden, *Visual Saliency Estimation by Nonlinearly Integrating Features Using Region Convariances*, Journal of Vision, vol.13, no.4, pp.1.20,2013.

[6] H.R. Tavakoli, E. Rahtu, and J. Heikkila, *Fast and Efficient Detection Using Sparse Sampling and Kernel Density Estimation*, Proc. of SCIA, pp.666–675,2011.

[7] Kageyama, and T. Sato, *The Yin-Yang grid:An Overset Grid in Spherical Geometry*, Geochem.Geophys., 5, Q09005, 2004.

[8] Bogdanova, A. Bur, and H. Hugli, *Visual Attention on The Sphere*, IEEE Trans. Image Process.,vol.17, no.11, 2008.

An Efficient Non-convex Mixture Method for Low-rank Tensor Completion

Shi Chengfei
School of Mechanical Sciences and engineering,
Huazhong University of Science and Technology
Wuhan, Hubei,China
scf@hust.edu.cn

Wan Li
School of Mechanical Sciences and engineering,
Huazhong University of Science and Technology
Wuhan, Hubei,China
wanli@hust.edu.cn

Huang Zhengdong
School of Mechanical Sciences and engineering,
Huazhong University of Science and Technology
Wuhan, Hubei,China
zdhuang@hust.edu.cn

Xiong Tifan
School of Mechanical Sciences and engineering,
Huazhong University of Science and Technology
Wuhan, Hubei,China
xiongtf@hust.edu.cn

ABSTRACT
For the problem of low-rank tensor completion, rank estimation plays an extremely important role. And among some outstanding researches, nuclear norm is often used as a substitute of rank in the optimization due to its convex property. However, recent advances show that some non-convex functions could approximate the rank better, which can significantly improve the precision of the algorithm. While, the complexity of non-convex functions also lead to much higher computation cost, especially in handling large scale matrices from the mode-n unfolding of a tensor. This paper proposes a mixture model for tensor completion by combining logDet function with Tucker decomposition to achieve a better performance in precision and a lower cost in computation as well. In the implementation of the method, alternating direction method of multipliers (ADMM) is employed to obtain the optimal tensor completion. Experiments on image restoration are carried out to validate the effective and efficiency of the method.

CCS Concepts
• Mathematics of computing →Approximation • Mathematics of computing → Nonconvex optimization • Computing methodologies →Image processing.

Keywords
Low-rank tensor completion; image restoration; logDet function; Tucker decomposition.

1. INTRODUCTION
Over last decades tensor completion, whose task is to estimate missing data via known data in a simple word, has been successfully applied in many fields such as signal processing [10], computer vision [19], numerical analysis [1], data mining [14], and so on. Recently, there has been an extensive utilization of tensor data expression in image inpainting [9], video inpainting [15], MRI data Recovery [18], and hyperspectral data recovery [11].

As a multi-dimensional generalization of matrix, tensor is more complex than matrix, especially for the rank definition which also plays an important role in tensor completion problem. With regard to tensor rank, there are two widely accepted definitions, CP-rank and *n*-rank, which are respectively developed from the CP and Tucker decompositions [2, 8]. Among the two types of rank, the evaluation of CP-rank turns out to be an NP-hard problem [4, 8], and therefore the existing research work more focuses on the *n*-rank. J. Liu et al. [13] proposed three traditionally methods for the problem of the *n*-rank based tensor completion: simple low-rank tensor completion (SiLRTC), fast low-rank tensor completion (FaLRTC), and high accuracy low-rank tensor completion (HaLRTC).

However, the solution schemes of the above methods all use N times of the singular value decomposition (SVD) at each iteration, where N is the dimension of a given tensor. And it would be much expensive when the data is on a large scale. Xu et al. [21] put forward a novel approach by applying the low-rank matrix factorizations to each mode matricization of the tensor. M. Filipovi et al. [2] employ Tucker decomposition to deal with this problem. What's more, some recent advances show that some non-convex functions like logDet function [6, 7] and weighted nuclear norm [3, 12] could approximate rank better, which have been proved to be more effective than the nuclear norm based methods. While the complexity of non-convex functions also lead to a much worse performance in computation cost, which is almost unacceptable.

Inspired by their research, this paper present a new mixture method for low-rank tensor completion with an objective to achieve a better performance in precision and a lower cost in computation as well, in which we combine the non-convex logDet function with Tucker decomposition.

The rest of the paper is organized as follows. First, some notations of tensors and related work are reviewed in **Sect. 2**. Afterwards, in **Sect. 3**, some theoretical issues about logDet function and Tucker decomposition are described in details, and then it comes up with

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301516

the tensor completion model based on the definitions. In **Sect. 4**, the alternating direction method of multipliers is adapted for solving the tensor completion model. After this, the experimental results are reported in **Sect. 5**. Finally, in **Sect. 6**, some conclusions are drawn from the research.

2. THEORETICAL BACKGROUND

2.1 Tensor Basics

This paper uses non-bold low-case script letters for scalars, e.g., x, bold low-case letters for vectors, e.g., **x**, non-bold upper-case letters for matrices, e.g., X, and bold upper-case script letters for tensors, e.g., $\boldsymbol{\mathcal{X}}$. An *N*th-order tensor is defined as $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, and $x_{i_1, i_2, \ldots i_N}$ is its $(i_1, \ldots, i_N)$-th component. Following definitions partially relate to the [17].

Definition 2.1.1 ***Inner product and Frobeninus norm:***

The inner product of two tensors $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, is defined as the sum of the products of their respective entries, i.e.,

$$< \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} > := \sum_{i_1}^{I_1} \sum_{i_2}^{I_2} \cdots \sum_{i_N}^{I_N} x_{i_1, i_2, \ldots i_N} \, y_{i_1, i_2, \ldots i_N}.$$

The Frobenius norm of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is the square root of the sum of the square of all its elements, i.e.,

$$\| \boldsymbol{\mathcal{X}} \|_{\mathrm{F}} = \sqrt{< \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} >}.$$

Definition 2.1.2 ***Unfolding and n-rank:***

The mode-*n* unfolding of a tensor X is denoted as $\boldsymbol{\mathcal{X}}_{(n)} = \mathrm{unfold}_n(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{I_n \times \prod_{i \neq n} I_i}$, which is a matrix with columns being the mode-*n* fibers of X in the lexicographical order. The inverse operator of unfolding is denoted as "fold", i.e., $\boldsymbol{\mathcal{X}} = \mathrm{fold}_n(\boldsymbol{\mathcal{X}}_{(n)})$.

The *n*-rank of an *N*th-order tensor $\boldsymbol{\mathcal{X}}$, denoted as $\mathrm{rank}_n(\boldsymbol{\mathcal{X}})$, is the rank of $\boldsymbol{\mathcal{X}}_{(n)}$, and the rank of X based on *n*-rank is defined as an array:

$$\mathrm{rank}(\boldsymbol{\mathcal{X}}) = (\mathrm{rank}(\boldsymbol{\mathcal{X}}_{(1)}), \ldots, \mathrm{rank}(\boldsymbol{\mathcal{X}}_{(N)})).$$

The tensor $\boldsymbol{\mathcal{X}}$ is low-rank, if $\boldsymbol{\mathcal{X}}_{(n)}$ is low-rank for all *n*.

Definition 2.1.3 ***Kronecker product:***

The Kronecker product of a matrix $\mathrm{A} \in \mathbb{R}^{I \times J}$ with a matrix $\mathrm{B} \in \mathbb{R}^{K \times L}$ is defined as

$$(\mathrm{A} \otimes \mathrm{B}) = \begin{bmatrix} a_{11}\mathrm{B} & \cdots & a_{1J}\mathrm{B} \\ \vdots & \ddots & \vdots \\ a_{I1}\mathrm{B} & \cdots & a_{IJ}\mathrm{B} \end{bmatrix} \in \mathbb{R}^{IK \times JL}$$

Definition 2.1.4 ***n-mode product*****:**

The n-mode (matrix) product of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ with a matrix $\mathrm{U} \in \mathbb{R}^{J \times I_N}$ is denoted by $(\boldsymbol{\mathcal{X}} \times_n \mathrm{U})$, a tensor $\in \mathbb{R}^{I_1 \times I_2 \times \ldots I_{\mathrm{n}-1} \times J \times I_{\mathrm{n}+1} \times \ldots \times I_N}$. And the element is defined as

$$(\boldsymbol{\mathcal{X}} \times_n \mathrm{U})_{i_1 \ldots i_{\mathrm{n}-1} j i_{\mathrm{n}+1} \ldots i_{\mathrm{N}}} = \sum_{i_n=1}^{I_N} x_{i_1 i_2 \ldots i_{\mathrm{N}}} u_{j i_{\mathrm{n}}}$$

Definition 2.1.5 ***Tucker decomposition*****:**

The Tucker decomposition [2] is a form of higher-order principal component analysis. It decomposes a tensor into a core tensor multiplied by a matrix along each mode. Thus, we have

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{G}} \times_1 \mathrm{A}^{(1)} \times_2 \mathrm{A}^{(2)} \times \cdots \times_N \mathrm{A}^{(\mathrm{N})}$$

where $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_N}$ and $\mathrm{A}^{(i)} \in \mathbb{R}^{I_i \times \mathrm{J}_i}$.

Furthermore, the mode-*n* unfolding of a tensor $\boldsymbol{\mathcal{X}}$ can be written as follow

$$\boldsymbol{\mathcal{X}}_{(n)} = \mathrm{A}^{(\mathrm{n})} \boldsymbol{\mathcal{G}}_{(n)} (\mathrm{A}^{(\mathrm{N})} \otimes \cdots \otimes \mathrm{A}^{(n+1)} \otimes \mathrm{A}^{(n-1)} \otimes \cdots \otimes \mathrm{A}^{(1)})$$

2.2 Tensor Completion

In this section, the problem of tensor completion is considered, aiming at recovering the missing entries using low-rank prior of the tensor, it is formulated as follow:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{X}}} \ \mathrm{rank}(\boldsymbol{\mathcal{X}}) \\ &s.t. \ \mathcal{P}_{\Omega}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{B}}. \end{aligned} \tag{2-1}$$

where $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ are N-mode tensors, Ω is the index set of observed entries, and $\mathcal{P}_{\Omega}(\cdot)$ is a sampling operator that keeps the entries in Ω and zeros out of others. However, the rank function $\mathrm{rank}(\cdot)$ is discrete and non-convex, and it is quiet tough to obtain a global solution to this problem. As mentioned above, the nuclear norm of matrix is used to approximate the rank of matrix. Based on the nuclear norm, the problem is converted into a convex optimization problem.

Then the convex substitution is extended to tensor by Liu et al. [13], and noticing that *n*-rank of a tensor is expressed as an array of mode-*n* rank, so the following convex problem can be solved as a variant of **(2-1)**:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{X}}} \ \| \boldsymbol{\mathcal{X}} \|_* = \sum_{i=1}^{N} \alpha_i \| \boldsymbol{\mathcal{X}}_{(i)} \|_* \\ &s.t. \ \mathcal{P}_{\Omega}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{B}}. \end{aligned}, \tag{2-2}$$

where α_i are constants satisfying $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$.

3. PROPOSED MIXTURE MODEL FOR TENSOR COMPLETION

The major reason for taking the non-convex logDet function into tensor completion, is that logDet function's approximation of rank(X) is much better than that of nuclear norm. The logDet-based matrix completion problem can be written as follows:

$$\begin{aligned} &\min_{X} \ logdet(X + I) \\ &s.t. \ \mathcal{P}_{\Omega}(X) = M \end{aligned}, \tag{3-1}$$

where $X \in \mathbb{R}^{n \times n}$, is a positive matrix, Ω is the set of locations corresponding to the observed entries, and I is an identity matrix. It is easy to indicate that $logdet(X + I) \leq \| X \|_*$.

As for the situation of tensor, following [6], the tensor completion problem can be represented as follows:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{X}}} \ \| \boldsymbol{\mathcal{X}} \|_* = \sum_{i=1}^{N} \alpha_i \mathrm{logdet}\left(\left(\boldsymbol{\mathcal{X}}_{(i)} \boldsymbol{\mathcal{X}}_{(i)}^{\mathrm{T}}\right)^{\frac{1}{2}} + I\right) \\ &s.t. \ \mathcal{P}_{\Omega}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{B}}. \end{aligned}, \tag{3-2}$$

where α_i are constants satisfying $\alpha_i \geq 0$, and $\sum_{i=1}^{N} \alpha_i = 1$.

Though the non-convex logDet function could significantly improve the precision of recover results, it also leads to a much higher computation cost, which is almost unacceptable. Thus, it's necessary to improve the computation efficiency.

According to the **Definition 2.1.5,** a tensor can be represented by a core tensor multiplied by a factor matrix along each mode. Notice that the size of the core tensor is generally smaller than the original tensor. By giving orthogonal constraint to each factor matrix, it can make sure that the rank of the core tensor is equal to

the original ones. Thus, a lower cost in computation can be achieved.

In this part, a mixture method of logDet function and Tucker decomposition for tensor completion is proposed. It's easy to get that:

$$\begin{aligned} \min_{C} \ & \sum_{i=1}^{N} \alpha_i \text{logdet}\left(\left(\boldsymbol{C}_{(i)}\boldsymbol{C}_{(i)}^{T}\right)^{\frac{1}{2}} + I\right) \\ s.t. \ & \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_{\mathbf{1}} \boldsymbol{U_1} \times_{\mathbf{2}} \cdots \times_{\boldsymbol{N}} \boldsymbol{U_N} \\ & U_i \in \text{St}(I_i, r_i), i = 1,2,\dots,N, \\ & \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{B}}. \end{aligned} \quad . \tag{3-3}$$

For convenient, we would define function $L(P)$ as follows:

$$\begin{aligned} L(P) &= \text{logdet}\left((PP^T)^{\frac{1}{2}} + I\right) \\ &= \log \textstyle\prod_{i=1}^{I_n}(\sigma_i(P) + 1). \\ &= \textstyle\sum_{i=1}^{I_n} \log(\sigma_i(P) + 1) \end{aligned} \tag{3-4}$$

In addition, we apply variable-splitting to V_i and introduce auxiliary variables $\boldsymbol{\mathcal{C}} = \boldsymbol{V}_i$, $i = 1,2,\dots,N$. What's more, in most application scenarios, the elements in tensor are non-negative, and thus the problem of **(3-3)** can be rewritten as follows:

$$\begin{aligned} & \min_{V_i} \ \textstyle\sum_{i=1}^{N} \alpha_i L\left(\boldsymbol{V}_{\boldsymbol{i},(i)}\right), \\ s.t. \ & \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_{\mathbf{1}} \boldsymbol{U_1} \times_{\mathbf{2}} \cdots \times_{\boldsymbol{N}} \boldsymbol{U_N}, U_i \in \text{St}(I_i, r_i) \\ & \boldsymbol{\mathcal{C}} = \boldsymbol{V_i}, i = 1,2,\dots,N, \\ & \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{X}} \in \mathbb{R}_{+}^{I_1 \times I_2 \times \dots \times I_N}. \end{aligned} \tag{3-5}$$

4. SOLUTION SCHEME

In this section, the numerical scheme for solving **(3-5)** is present. First, we define the simplex constraint set by $\Delta \coloneqq \{\boldsymbol{\mathcal{X}} : \mathcal{P}_{\boldsymbol{\Omega}}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{X}} \geq 0\}$. We keep the constraints and deal with them through the solution while solving $\boldsymbol{\mathcal{X}}$.

The partial augmented Lagrangian formulation for **(3-5)** is then:

$$\mathcal{L}(U_i, V_i, \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{X}} \in \Delta, \boldsymbol{\theta}_i) =$$

$$\textstyle\sum_{i=1}^{N}(\alpha_i L\left(\boldsymbol{V}_{i,(i)}\right) + \left(\frac{u}{2} \parallel \boldsymbol{\mathcal{C}} - \boldsymbol{V}_i \parallel_F^2 + \langle \boldsymbol{\theta}_i, \boldsymbol{\mathcal{C}} - \boldsymbol{V}_i \rangle\right)) + \frac{\lambda}{2} \parallel \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{C}} \times_{\mathbf{1}} \boldsymbol{U_1} \times_{\mathbf{2}} \cdots \times_{\boldsymbol{N}} \boldsymbol{U_N} \parallel_F^2,$$

where μ,λ >0 are penalty parameters, and $\boldsymbol{\theta_i} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is an estimate of the lagrange multiplier.

Then the $\mathcal{L}(U_i, \boldsymbol{V_i}, \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{X}} \in \Delta, \boldsymbol{\theta_i})$ is minimized with respect to one of the variables while the others are held fixed at each iteration, and the sub problems are as follows:

$$\begin{cases} Step\ 1\text{: } \mathrm{U}_i^{\mathrm{k+1}} \leftarrow \arg\min \mathcal{L}\left(U_i, \boldsymbol{V}_i^k, \boldsymbol{\mathcal{C}^k}, \boldsymbol{\mathcal{X}^k} \in \Delta, \boldsymbol{\theta}_i^k\right) \\ Step\ 2\text{: } \boldsymbol{V}_i^{\mathrm{k+1}} \leftarrow \arg\min \mathcal{L}\left(U_i^{k+1}, \boldsymbol{V}_i, \boldsymbol{\mathcal{C}^k}, \boldsymbol{\mathcal{X}^k} \in \Delta, \boldsymbol{\theta}_i^k\right) \\ Step\ 3\text{: } \boldsymbol{\mathcal{C}^{k+1}} \leftarrow \arg\min \mathcal{L}\left(U_i^{k+1}, \boldsymbol{V}_i^{k+1}, \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{X}^k} \in \Delta, \boldsymbol{\theta}_i^k\right) \\ Step\ 4\text{: } \boldsymbol{\mathcal{X}}^{\mathrm{k+1}} \leftarrow \arg\min \mathcal{L}\left(U_i^{k+1}, \boldsymbol{V}_i^{k+1}, \boldsymbol{\mathcal{C}^{k+1}}, \boldsymbol{\mathcal{X}} \in \Delta, \boldsymbol{\theta}_i^k\right) \\ Step\ 5\text{: } \boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + u\, \partial \mathcal{L}\left(U_i^{k+1}, \boldsymbol{V}_i^{k+1}, \boldsymbol{\mathcal{C}^{k+1}}, \boldsymbol{\mathcal{X}}^{\mathrm{k+1}} \in \Delta, \boldsymbol{\theta}_i\right) / \partial\, \boldsymbol{\theta}_i^k \end{cases}$$

A. Updating U_i^{k+1}

In Step 1, we solve the following sub problem:

$$U_i^{k+1} = \underset{U_i}{\text{argmin}} \left\{ \begin{matrix} \frac{\lambda}{2} \parallel \boldsymbol{\mathcal{X}^k} - \boldsymbol{\mathcal{C}^k} \times_{\mathbf{1}} \boldsymbol{U_1^k} \times_{\mathbf{2}} \cdots \times_{\boldsymbol{N}} \boldsymbol{U_N^k} \parallel_F^2 \\ s.t.\, U_i \in \text{St}(I_i, r_i) \end{matrix} \right\}. \tag{4-1}$$

For this sub-problem, it is a special Tucker decomposition problem, which is actually called HOOI. And research [16] gave the solution scheme, in which U_i is solved iteratively. However, it is obviously not a proper way to solve it iteratively in our method since it will cost much more time. Thus, an inaccurate iterative strategy is adopted, just updating U_i once. Therefore, it comes with

$$\begin{aligned} W = \boldsymbol{\mathcal{X}^k} \times_{\mathbf{1}} \left(\boldsymbol{U_1^{k+1}}\right)^{\boldsymbol{T}} \times_{\mathbf{2}} \cdots \times_{\boldsymbol{i-1}} \left(\boldsymbol{U_{i-1}^{k+1}}\right)^{\boldsymbol{T}} \times_{\boldsymbol{i+1}} \left(\boldsymbol{U_{i+1}^{k+1}}\right)^{\boldsymbol{T}} \cdots \\ \times_{\boldsymbol{N}} \left(\boldsymbol{U_N^{k+1}}\right)^{\boldsymbol{T}}. \end{aligned} \tag{4-2}$$

$$U_i^{k+1} = \boldsymbol{SVD}(r_i, W_{(i)}). \tag{4-3}$$

B. Updating $\boldsymbol{V_i^{k+1}}$

In Step 2, we solve the following sub problem:

$$\boldsymbol{V_i^{k+1}} = \underset{\boldsymbol{V_i}}{\text{argmin}} \left\{ \alpha_i L\left(\boldsymbol{V}_{\boldsymbol{i},(i)}\right) + \left(\frac{u}{2} \parallel \boldsymbol{\mathcal{C}} - \boldsymbol{V}_i \parallel_F^2 + \langle \boldsymbol{\theta_i}, \boldsymbol{\mathcal{C}} - \boldsymbol{V_i} \rangle\right) \right\},$$

Merging the last two terms, obtain:

$$\boldsymbol{V_i^{k+1}} = \underset{\boldsymbol{V_i}}{\text{argmin}} \left\{ \alpha_i L\left(\boldsymbol{V}_{\boldsymbol{i},(i)}\right) + \frac{u}{2} \parallel \boldsymbol{\mathcal{C}} - \boldsymbol{V_i} + \frac{\boldsymbol{\theta_i}}{u} \parallel_F^2 \right\}. \tag{4-5}$$

Before solving (4-10), it is necessary to introduce **Lemma 1**.

Lemma 1: *Given $Z \epsilon \mathbb{R}^{m \times n}, A \epsilon \mathbb{R}^{m \times n}, u > 0$, $F(Z)$ is a unitarily invariant function .Then the minimization problem is as below:*

$$Z^* = \underset{Z}{arg\,min} F(Z) + \frac{u}{2} \parallel Z - A \parallel_F^2, \tag{4-6}$$

which is equivalent to :

$$\sigma^* = \underset{Z}{arg\,min} f(\sigma) + \frac{u}{2} \parallel \sigma - \sigma_A \parallel_F^2, \tag{4-7}$$

so that $Z^* = Udiag(\sigma^*)V^T$, with the SVD of A being $Udiag(\sigma_A)V^T$.

According to the first-order optimality condition, the gradient of the objective function of **(4-5)** with respect to each singular value should vanish. For the logDet function, we have

$$\sigma_n^* \in \{\tfrac{1}{1+\sigma_n} + \tfrac{u}{\alpha_i}\left(\sigma_n - \sigma_{n,C}\right) = 0 \ s.t. \ \sigma_n \geq 0\}, \tag{4-8}$$

where $\sigma_{n,C}$ is the singular value of $(\boldsymbol{C}_{(i)} + \frac{\theta_{i,(i)}}{u})$. The equation of (4-8) is quadratic and has two roots. If $\sigma_{n,C} = 0$, the minimizer σ_n^* will be 0; otherwise, there must exist a unique minimizer.

Finally, we can update the $\boldsymbol{V_i^{k+1}}$ with

$$\boldsymbol{V_{i,(i)}^{k+1}} = Udiag(\sigma^*)V^T, \tag{4-9}$$

where $\sigma^* = \{\sigma_1^*, \sigma_2^*, \dots, \sigma_n^*\}$, $\sigma_C = \{\sigma_{1,C}, \sigma_{2,C}, \dots, \sigma_{n,C}\}$. Here, σ_n^* and σ_C respectively are the singular values of $\boldsymbol{V_{i,(i)}^{k+1}}$ *and* $(\boldsymbol{C}_{(i)} + \frac{\theta_{i,(i)}}{u})$, and $(\boldsymbol{C}_{(i)} + \frac{\theta_{i,(i)}}{u}) = Udiag(\sigma_C)V^T$.

C. Updating $\boldsymbol{\mathcal{C}^{k+1}}$

In Step 3, we solve the following sub problem:

$$\begin{aligned} \boldsymbol{\mathcal{C}^{k+1}} = \underset{\boldsymbol{\mathcal{C}}}{\text{argmin}} \left\{ \frac{\lambda}{2} \parallel \boldsymbol{\mathcal{X}^k} - \boldsymbol{\mathcal{C}} \times_{\mathbf{1}} \boldsymbol{U_1^{k+1}} \times_{\mathbf{2}} \cdots \times_{\boldsymbol{N}} \boldsymbol{U_N^{k+1}} \parallel_F^2 + \right. \\ \left. \textstyle\sum_{i=1}^{N} \frac{u}{2} \parallel \boldsymbol{\mathcal{C}} - \boldsymbol{V}_i^{k+1} + \frac{\boldsymbol{\theta}_i^k}{u} \parallel_F^2 \right\}. \end{aligned} \tag{4-10}$$

It is evident that the objective function of **(4-10)** is a strongly convex quadratic function, which can be solved directly by making its first derivative to zero. Therefore, it comes with:

$$\frac{\mathcal{L}\left(U_i^{k+1}, \boldsymbol{V_i^{k+1}}, \boldsymbol{C}, \boldsymbol{\mathcal{X}^{k}} \in \Delta, \boldsymbol{\theta_i^k}\right)}{\partial \mathbf{C}} = \lambda\left(C - \boldsymbol{\mathcal{X}^k} \times_1 \left(\boldsymbol{U_1^{k+1}}\right)^T \times_2 \cdots \times_N \left(\boldsymbol{U_N^{k+1}}\right)^T\right) + \sum_{i=1}^{N} u(\boldsymbol{C} - \boldsymbol{V_i^{k+1}} + \frac{\boldsymbol{\theta_i^k}}{u}) = 0, \quad (4\text{-}11)$$

The equation **(4-11)** can be solved easily, then the optimization result of $\boldsymbol{C^{k+1}}$ is as follow**:**

$$\boldsymbol{C^{k+1}} = \begin{pmatrix} \lambda \boldsymbol{\mathcal{X}^k} \times_1 \left(\boldsymbol{U_1^{k+1}}\right)^T \times_2 \cdots \times_N \left(\boldsymbol{U_N^{k+1}}\right)^T \\ + \sum_{i=1}^{N} u\left(\boldsymbol{V_i^{k+1}} - \frac{\boldsymbol{\theta_i^k}}{u}\right) \end{pmatrix} \Big/ (\boldsymbol{N}\mu + \lambda). \quad (4\text{-}12)$$

D. Updating $\mathcal{X}^{k+1}$

In Step 4, we solve the following sub problem:

$$\boldsymbol{\mathcal{X}^{k+1}} = \underset{\boldsymbol{\mathcal{X}}}{\operatorname{argmin}} \left\{\frac{\lambda}{2} \parallel \boldsymbol{\mathcal{X}} - \boldsymbol{C^{k+1}} \times_1 \boldsymbol{U_1^{k+1}} \times_2 \cdots \times_N \boldsymbol{U_N^{k+1}} \parallel_F^2\right\}. \quad (4\text{-}13)$$

Dealing with this problem, simplex constraint Δ should be taken into consideration:

$$\boldsymbol{\mathcal{X}^{k+1}} = \begin{cases} \boldsymbol{C^{k+1}} \times_1 \boldsymbol{U_1^{k+1}} \times_2 \cdots \times_N \boldsymbol{U_N^{k+1}}, & x_{i_1, i_2, \ldots i_N} \notin \boldsymbol{\Omega} \\ \boldsymbol{\mathcal{B}}, & x_{i_1, i_2, \ldots i_N} \in \boldsymbol{\Omega} \end{cases}. \quad (4\text{-}14)$$

E. Updating $\boldsymbol{\theta_i^{k+1}}$

In Step 5, we solve the following sub problem:

$$\boldsymbol{\theta_i^{k+1}} = \theta_i^k + u\, \partial \mathcal{L}\left(U_i^{k+1}, \boldsymbol{V_i^{k+1}}, \boldsymbol{C^{k+1}}, \boldsymbol{\mathcal{X}^{k+1}} \in \Delta, \theta_i\right) / \partial\, \theta_i^k. \quad (4\text{-}15)$$

It is apparent that:

$$\boldsymbol{\theta_i^{k+1}} = \theta_i^k + u(\boldsymbol{C^{k+1}} - \boldsymbol{V_i^{k+1}}). \quad (4\text{-}16)$$

Based on the previous analysis, an iterative scheme for tensor completion, as outlined in **Algorithm 1,** is developed. And **Algorithm 1** can be accelerated by adaptively changing u. In this paper, a multiple $t = 1.05$ is set to increase u iteratively, i.e., $u^{k+1} = tu^k$. However, the proposed method is non-convex, and the convergence properties of the ADMM in theory is still an open issue [5].

Algorithm 1 Mixture of logdet function and Tucker decomposition

Input: the tensor B, the index set Ω, upper bounds $(r_1, r_2, ..., r_N)$, the parameters μ, λ, Maxiter, Tol and t

Initialization: set $C = rand(r_1, r_2, ..., r_N)$, $U_i = rand(I_i, r_i)$, $V_i = C$, $i \epsilon \{1, ..., N\}$

for k=1 to Maxiter **do**
  for i=1 to N **do**
    Update U_i^{k+1} By (4-3);
  end for
  for i=1 to N **do**
    Update V_i^{k+1} By (4-9);
  end for
  Update C^{k+1} By (4-12);
  Update $\mathcal{X}^{k+1}$ By (4-14);
  for n=1 to N **do**
    Update θ_i^{k+1} By (4-16);
  end for
  Check the convergence condition $\|\mathcal{X}^{k+1} - \mathcal{X}^k\|_F / \|\mathcal{X}^k\|_F < Tol$
end for

Output $\mathcal{X}$

5. EXPERIMENT

In this section, the effectiveness and efficiency of the proposed LTLRTC approach for tensor completion are evaluated on real-world data. And we compare our method with two other state-of-the-art methods: HaLRTC [13] and LogDet [6]. Among them, HaLRTC is a classical nuclear norm based method, which has highly accuracy and is also efficiency, and LogDet is a non-convex based method which has also revealed the great performance in accuracy on low-rank tensor completion.

All the tests are performed under Windows 10 and Matlab version 9.0.0.341360 (R2016a) with an Intel(R) Core(TM) i5-3230M CPU at 2.60GHz and 4GB of memory.

The test color videos are in the YUV format, we choose the video named "akiyo" and transform the video into a fourth-order tensor, after that we employ the first 100 frames of the video so that the size of the tensor is $144 \times 176 \times 3 \times 100$.

Furthermore, as to the parameters, we set the weights $\alpha_i = I_i / \sum_{i=1}^{N} I_i$, the condition of convergence $\mathrm{Tol} = 10^{-4}$ and the maximum number of iteration $\mathrm{Maxiter} = 500$. The parameters u, λ for our method are $\mathrm{u} = 1\mathrm{e}^{-8}, \lambda = 10$. The parameter rho for HaLRTC is set to be $rho = 1\mathrm{e}^{-6}$. And the parameters β, ε for logDet are $\beta = 1\mathrm{e}^{-8}, \varepsilon = 1\mathrm{e}^{-5}$ according to [6]. For our method, the upper bounds, the size of the core tensor, is approximated by the numbers of singular values which is larger than the 1% of the largest one. Thus, the estimation upper bounds is $\mathrm{r} = (60,60,3,20)$.

And following [20], this paper further quantify the quality of the tensor completion via peak signal-to-noise ratio (PSNR):

$$\mathrm{PSNR}(\mathcal{X}, \mathrm{T}) = 10 \log_{10} \frac{n \boldsymbol{T}_{max}^2}{\parallel \boldsymbol{\mathcal{X}} - \boldsymbol{T} \parallel_F^2},$$

where $\boldsymbol{\mathcal{X}}$ is the estimated tensor, $\boldsymbol{T}$ is the original tensor, n denotes the total number of the pixels in the tensor, and $\boldsymbol{T}_{max}^2$ represents the maximum pixel value of the original tensor.

Table 1. Results (RSE(10^{-2}) and Times(s)) for different methods with different sampling rates.

SR	HaLRTC		LogDet		Ourmethod	
	RSE	Time	RSE	Time	RSE	Time
20%	9.06	184	4.78	930	3.99	441
30%	6.62	176	3.49	839	3.42	211
40%	4.99	159	2.77	721	3.06	153
50%	3.77	186	2.14	830	2.74	96.8
60%	2.84	167	1.63	808	2.43	93.9
70%	2.08	184	1.22	841	2.09	71.5
80%	1.45	159	0.89	850	1.70	50.9

From **Table 1** and **Figure 1**, the following findings can be reached: (1) In general, LogDet gets the best results in accuracy for all sampling rates except 20% and 30%, however, the computation cost is much higher than any other method, which is

almost unacceptable; (2) Compared with HaLRTC, which is also both effective and efficiency, our method performs better in recover quality when the sampling rate is lower than 70%, and the time cost by our method is lower when the sampling rate is over 40%; (3) Considering both the accuracy and computation efficiency of the results, our method is superior to LogDet and HaLRTC.

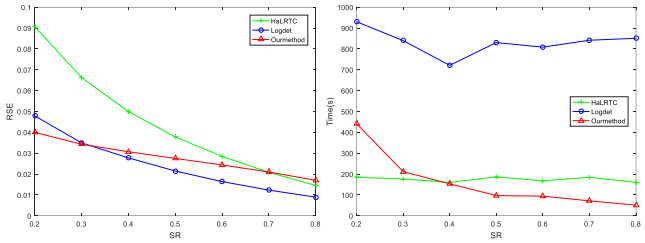


Figure 1. RSE and time cost results for different methods with different sampling rates from 20% to 80%.

Since the time consumption of LogDet is unacceptable, and the gap of the performance on recover accuracy between LogDet and our method is narrow. This paper mainly focus on the comparison between HaLRTC and our method. Thus, for a more intuitively view on the recovered results, a specific frame of the recovered video is displayed in **Figure 2**. What's more, for the sake of probing into the effect of this two methods more carefully and deeply, we evaluate the PSNR values for every frame with different sampling rates. The results are shown in **Figure 3**. It's easy to see that our method performs better and as the growing of sampling rate, the gap is narrowing. Moreover, more videos are also tested, and the results are summarized in **Table 2**.

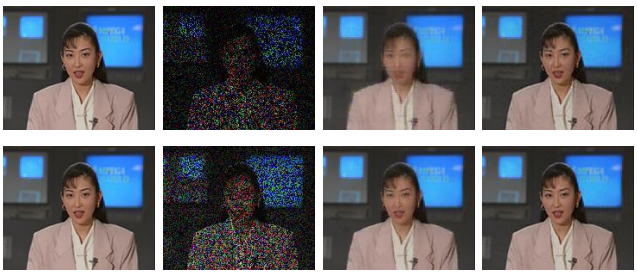


Figure 2. The recovered results (the 60th frame) of the video. From left to right: the original frame, masked frame, and the recovered results by HaLRTC and our Method. From top to bottom: the sampling rates are 20% and 40%.

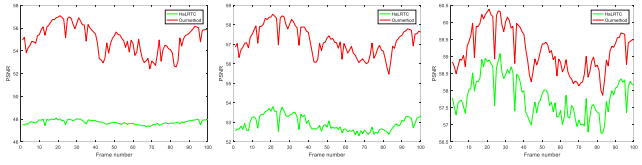


Figure 3. The PSNR values for each frame and the sampling rates are 20%, 40% and 60%.

6. CONCLUSION

In this paper, a mixture model for low-rank tensor completion by integrating the logDet function with the Tucker decomposition is proposed. In addition, an efficient iterative scheme is employed to deal with the proposed model. Finally, the experimental results on real-world data demonstrate both the effectiveness and practicability of our method. However, this paper finds that the value of upper bounds has great effect on the recover results, while, it's difficult to accurately select the appropriate value, and actually, the problem of rank approximation in practical application is still an open issue. In the future, we will further explore this problem.

Table 2 Results comparison for different videos

Video	SR	HaLRTC		Our method		
		RSE	Time	RSE	Time	Bounds
Hall	0.3	7.34	**135**	**4.09**	201	(60,60,3,20)
	0.5	4.40	161	**3.29**	**114**	
	0.7	2.62	159	**2.51**	**45.2**	
Bridge	0.3	4.54	132	**3.01**	**101**	(50,50,3,10)
	0.5	3.13	133	**2.50**	**65.4**	
	0.7	2.09	124	**1.93**	**49.4**	
Claire	0.3	4.93	180	**2.48**	**156**	(50,50,3,20)
	0.5	2.77	177	**1.97**	**84.0**	
	0.7	1.58	170	**1.50**	**37.8**	
Foreman	0.3	8.07	**117**	**4.22**	474	(70,70,3,40)
	0.5	4.67	**114**	**3.13**	133	
	0.7	2.54	127	**2.33**	**90.1**	
Highway	0.3	4.06	**136**	**2.75**	146	(50,50,3,20)
	0.5	2.71	150	**2.23**	**77.5**	
	0.7	1.77	142	**1.71**	**43.4**	
Suzie	0.3	7.00	**146**	**4.29**	609	(50,50,3,40)
	0.5	4.21	152	**3.33**	**145**	
	0.7	**2.50**	**146**	2.52	152	
Container	0.3	7.19	**136**	**4.08**	313	(60,60,3,20)
	0.5	4.28	139	**3.04**	**128**	
	0.7	2.43	133	**2.26**	**91.5**	
mother-daughter	0.3	5.88	**142**	**3.45**	167	(60,60,3,20)
	0.5	3.51	156	**2.81**	**104**	
	0.7	**1.98**	151	2.15	**44.9**	
Silent	0.3	6.82	258	**4.03**	**245**	(60,60,3,30)
	0.5	3.96	**141**	**3.22**	181	
	0.7	**2.22**	173	2.45	**66.5**	

7. ACKNOWLEDGEMENTS

we would like to express great thankfulness to the reviewers for their helpful suggestions. We also appreciate J.Liu for sharing the codes of HaLRTC algorithm.

8. REFERENCES

[1] L G Beylkin and M. J. Mohlenkamp. 2002. Numerical operator calculus in higher dimensions. *Proceedings of the National Academy of Sciences of the United States of America* 99, 16 (2002),10246–10251.

[2] Filipovi, Marko, Juki, and Ante. 2015. *Tucker factorization withmissing data with application to low-n-rank tensor completion*. Kluwer Academic Publishers. 677–692 pages.

[3] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng.2014. Weighted Nuclear Norm Minimization with Application to Image Denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2862–2869.

[4] H and Johan Aring. 1990. *Tensor rank is NP-complete*. Academic Press, Inc. 644–654 pages.

[5] Bingsheng He, Min Tao, and Xiaoming Yuan. 2012. Alternating Direction Method with Gaussian Back Substitution for Separable Convex Programming. *Siam Journal on Optimization* 22, 2 (2012), 313–340.

[6] Teng Yu Ji, Ting Zhu Huang, Xi Le Zhao, Tian Hui Ma, and Liang Jian Deng. 2017. A non-convex tensor rank approximation for tensor completion. *Applied Mathematical Modelling* 48 (2017), 410–422.

[7] Zhao Kang, Chong Peng, and Qiang Cheng. 2015. Robust Subspace Clustering via Smoothed Rank Approximation. *IEEE Signal Processing Letters* 22, 11 (2015), 2088–2092.

[8] Tamara G. Kolda and Brett W. Bader. 2009. Tensor Decompositions and Applications. *Siam Review* 51, 3 (2009), 455–500.

[9] N Komodakis. 2006. Image Completion Using Global Optimization. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. 442–452.

[10] Lieven De Lathauwer and Joos Vandewalle. 2004. Dimensionality reduction in higher-order signal processing and rank-(R 1 , R 2 ,, R N) reduction in multilinear algebra. *Linear Algebra and Its Applications* 391, 1 (2004), 31–55.

[11] Nan Li and Baoxin Li. 2015. Tensor completion for on-board compression of hyperspectral images. In *IEEE International Conference on Image Processing*. 517–520.

[12] Yu Fan Li, Yan Jiao Zhang, and Zheng Hai Huang. 2014. A reweighted nuclear norm minimization algorithm for low rank matrix recovery. *J. Comput. Appl. Math.* 263, C (2014), 338–350.

[13] J. Liu, P Musialski, P Wonka, and J. Ye. 2009. Tensor completion for estimating missing values in visual data.. In *IEEE International Conference on Computer Vision*. 2114–2121.

[14] Morten M?rup. 2011. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery* (2011), 24–40.

[15] K. A Patwardhan, G Sapiro, and M Bertalmio. 2007. Video Inpainting Under Constrained Camera Motion. *IEEE Trans Image Process* 16, 2 (2007), 545–553.

[16] Bernard N. Sheehan and Yousef Saad. 2007. Higher Order Orthogonal Iteration of Tensors (HOOI) and its Relation to PCA and GLRAM. In *Siam International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, Usa*.

[17] Ledyard R Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31, 3 (1966), 279–311.

[18] V. Nivitha Varghees, M. Sabarimalai Manikandan, and Rolant Gini. 2012. Adaptive MRI image denoising using total-variation and local noise estimation. In *International Conference on Advances in Engineering, Science and Management*. 506–511.

[19] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. 2005. Face transfer with multilinear models. *Acm Transactions on Graphics* 24, 3 (2005), 426–433.

[20] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13, 4 (2004), 600–612.

[21] Yangyang Xu, Ruru Hao, Wotao Yin, and Zhixun Su. 2017. Parallel matrix factorization for low-rank tensor completion. *Inverse Problems and Imaging* 9, 2 (2017), 601–624.

Clustering color segmentation in multi-color space

Ting Tu
School of Mechanical and Automotive Engineering
Shanghai University of Engineering Science
Shanghai 201620，China
+86-18621030693
985911610@qq.com

Zhifeng Zhou*
School of Mechanical and Automotive Engineering
Shanghai University of Engineering Science
Shanghai 201620，China
+86-13564125476
zhousjtu@126.com

Peng Xiao
School of Mechanical and Automotive Engineering
Shanghai University of Engineering Science
Shanghai 201620，China
+86-13161830451
439992880@qq.com

ABSTRACT
The quality of image segmentation seriously affects the results of digital images analysis and calculation. Aiming at the problem that the parameters of classification and initial center need to be entered for color image segmentation by using k-means clustering algorithm, a k-means clustering algorithm based on multiple color space is proposed for color image segmentation. Firstly, the study of color space and its mutual transformation showed that HIS color space and CIELAB color space were more consistent with human visual characteristics. Then, the gaussian mixture model was used to solve the H component of HIS color space. The gaussian mixture model parameters were obtained by the maximum expectation algorithm, and the clustering number and the initial center were obtained by optimizing the model with the Akaike information criterion. The simulated annealing algorithm is used to improve the traditional k-means algorithm, which avoids the problem that the traditional k-means algorithm is easy to fall into the local optimum. Then, k-means clustering analysis was conducted on (a, b) two-dimensional components under CIELAB color space, and finally color segmentation was conducted by the clustering results. The experiment on color segmentation of flower cluster images in MATLAB. Compared with the traditional k-means clustering algorithm, the proposed algorithm has good effect of color segmentation without human intervention and has good robustness.

CCS Concepts
• **Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems → Image segmentation**

Keywords
color segmentation; color space; k-means clustering; gaussian mixture model.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301528

1. INTRODUCTION
With the development of computer and machine vision, digital image processing has been widely used in aerospace, aviation, medicine, manufacturing and other fields due to its advantages of non-contact, fast, efficient, and high reliability. Color images are widely used because they contain more information than grayscale images. Color segmentation is one of the difficulties in color image processing. How to segment color images effectively and quickly has been paid more attention by researchers in national or aboard. In the CIELAB color space, researchers such as Amruta B Patil used the OTSU algorithm to segment the flower image [1], and applied the optimal threshold to three components of L, a and b. And the method can extract the shape, color and texture better; the literature [2] proposes a color image segmentation algorithm in unsupervised, which integrates the color distance into the pulse-coupled neural network by connecting the control unit. Image segmentation is automatically performed by iteration. In recent years, color segmentation algorithm based on clustering has been widely used. This algorithm mainly uses the color feature information of each pixel to cluster according to the similarity of each pixel property. Each pixel is labeled in a different category, and independent of each other, It isn't involved the information about the pixel position [3]. After Siti Noraini Sulaiman and Nor Ashidi Mat Isa proposed adaptive fuzzy k-means color segmentation algorithm [3], the literature [4-7] proposed an improved image segmentation algorithm based on clustering analysis for segmentation time and effect.

For the problem that the parameters of classification and initial center need to be entered when using k-means clustering algorithm. This paper proposes a k-means algorithm in multi-color space to color segment the color image. Firstly, the study of color space and its conversion relationship showed that the HIS color space and the CIELAB color space are more consistent with human visual characteristics. Then the maximum expected algorithm is used to solve the Gaussian mixture model of the H component under the HIS model, and the clustering number and the initial center were obtained by optimizing the model with the Akaike information criterion. The simulated annealing algorithm is used to improve the traditional k-means algorithm, which avoids the problem that the traditional k-means algorithm is easy to fall into the local optimum. Then, k-means clustering analysis was conducted on (a, b) two-dimensional components under CIELAB color space, and finally color segmentation was conducted by the clustering results.

2. COLOR SPACE

The color space is a description of the color under certain standards to visualize the color. The color space is represented by a three-dimensional model that uses three-dimensional coordinates to represent the three component parameters of the color space.

2.1 HIS Color Space

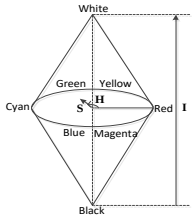


Figure 1. Model diagram of HIS color space.

HIS color space describes colors by H (Hue), I (Intensity) and S (Saturation). Hue indicates the color category, and saturation is the degree of color shade. The HIS color space is described by a biconical space model. As shown in Figure 1, the model is complex, but it can show the changes of the three components. The HIS color space has two important features: (1) the I component is independent of color information; (2) the H and S components are linked closely with the mode of perception [8]. At the same time, the components of the HIS color space are independent of each other and can be processed separately, which greatly reduces the workload and complexity of the processing.

The conversion formula from RGB color space to HIS color space is formula (1). It can be seen from equation (1) that when I is zero, S is meaningless; if S is zero, H is meaningless. Therefore, when converting colors from RGB to HIS, the H component produces a singular point that cannot be eliminated [8], and the pixels with lower saturation are ignored.

$$\begin{aligned} H &= \begin{cases} \theta & \text{if } G \ge B \\ 2\pi - \theta & \text{if } G < B \end{cases}, \quad with\ \theta = \cos^{-1}\left(\frac{(R-G)+(R-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}} \right) \\ I &= \frac{R+G+B}{3} \\ S &= 1 - \frac{\min(R,G,B)}{I} \end{aligned} \tag{1}$$

2.2 CIELAB Color Space

The CIELAB color space is an international standard which developed by the International Lighting Association CIE, and it is a device-independent color space. The CIELAB is considered to have the best uniformity except that the blue region needs to be converted to hue Angle [9,10]. The CIELAB uses L, a and b to represent the color: L represents illuminance from dark to bright, $L \in [0,100]$; a represents a change from green to red, $a \in [-127,128]$; and b represents a change from yellow to blue, $b \in [-128,127]$. The CIELAB color space has a wider color gamut that is closer to human visual perception.

Converting colors from RGB to CIELAB first needs to be converted to XYZ space, as shown in equation (2), and then converted from XYZ space to CIELAB color space, as shown in equation (3).

$$\begin{cases} X = 0.49 \times R + 0.31 \times G + 0.2 \times B \\ Y = 0.177 \times R + 0.812 \times G + 0.011 \times B \\ Z = 0.01 \times G + 0.99 \times B \end{cases} \tag{2}$$

$$\begin{cases} L = 116 f(Y) - 16 \\ a = 500\left[f\left(\frac{X}{0.982}\right) - f(Y) \right] \\ b = 200\left[f(Y) - f\left(\frac{Z}{1.183}\right) \right] \end{cases} \tag{3}$$

with:

$$f(X) = \begin{cases} 7.787X + 0.138, X \le 0.008856 \\ X^{\frac{1}{3}}, X > 0.008856 \end{cases} \tag{4}$$

3. GAUSSIAN MIXTURE MODEL

3.1 Gaussian Mixture model

The Gaussian mixture model (GMM) is a method that decomposes a thing into several models based on Gaussian functions, and it is widely used in data analysis and pattern recognition [11]. Considering that the histogram data of the H component in the HIS color space are all positive and there are multiple peaks, it is modeled as a Gaussian mixture model, and each component is modeled as a peak.

Let the histogram data of the H component be $X_n = (x_1, x_2, \cdots, x_n)$, then the Gaussian mixture model is defined as:

$$p(x|\theta) = \sum_{j=1}^{k} \tau_j g(x|\theta_j), x \ge 0 \tag{5}$$

Where k is the number of components in the GMM, τ_j is the probability of being selected for class *j*, then $\theta_j = (\mu_j, \sigma_j^2)$,with μ_j is the mean of class *j*, σ_j^2 is the standard deviation of class *j*, $g(x|\theta_j)$ is the Gaussian distribution density, as follows:

$$g(x|\theta_j) = g(x|u_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2} \right) \tag{6}$$

3.2 Maximum Expectation Algorithm

The log-likelihood function is used to solve the parameters of the Gaussian mixture model. The log-likelihood function is given by equation (7):

$$L(\theta | X_n) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \tau_j g(x_i | \mu_j, \sigma_j^2) \right) \tag{7}$$

It is necessary to introduce a hidden variable of $Z = \{z_i\}$ to describe whether the element object of the data belongs to the Gaussian mixture model by using the EM algorithm. Where $z_i = j$ shows that x_i is belonging to the GMM component of *j*, and it is unknown.

The steps to solve the GMM parameters using the EM algorithm are as follows:

(1) Initialize the parameters of Gaussian mixture model, let $\tau_j = 1/k$, and obtain the parameters of μ_j 、 σ_j^2 by closed estimate based on moment method. The first moment and the second moment are μ_j and σ_j^2 ;

(2) Given the estimated parameters of θ^m after m iterations, the weights belonging to component *j* in the data are $p(Z_i = j | x_i, \theta^m)$:

$$p(Z_i = j \mid x_i, \theta^m) = \frac{p(Z_i = j \mid \theta^m) g(x_i \mid \theta_j^m)}{\sum_{t=1}^{k} p(Z_i = t \mid \theta^m) g(x_i \mid \theta_t^m)} = \frac{\tau_j g(x_i \mid \theta_j^m)}{\sum_{t=1}^{k} \tau_t g(x_i \mid \theta_t^m)} \quad (8)$$

Then the expectation of the log-likelihood function for the hidden variable is

$$\begin{aligned} Q\left(\theta \mid \theta^m, X_n\right) &= E_{Z|\theta^m, X_n}\{L(\theta \mid X_n, Z)\} \\ &= \sum_{i=1}^{n} E_{Z|\theta^m, x_i}\{\log g(x_i \mid \theta_{z_i}, Z_i = z_i) + \log p(Z_i = z_i \mid \theta)\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} p(Z_i = j \mid x_i, \theta^m)(\log g(x_i \mid \theta_j) + \log \tau_j) \end{aligned} \quad (9)$$

(3) Maximize the expectation of the log-likelihood function and obtain new estimation parameters for the model by

$$\theta^{m+1} = \arg\max_{\theta} Q(\theta \mid \theta^m, X_n) \quad (10)$$

(4) The GMM parameter are

$$\tau_j = \frac{1}{n}\sum_{i=1}^{n} p(Z_i = j \mid x_i, \theta^m) \quad (11)$$

$$\log(\alpha_j) - \psi(\alpha_j) = \log\left(\frac{\sum_{i=1}^{n} x_i p(Z_i = j \mid x_i, \theta^m)}{\sum_{i=1}^{n} p(Z_i = j \mid x_i, \theta^m)}\right) - \left(\frac{\sum_{i=1}^{n} p(Z_i = j \mid x_i, \theta^m) \log x_i}{\sum_{i=1}^{n} p(Z_i = j \mid x_i, \theta^m)}\right) \quad (12)$$

$$\psi(\alpha) = \frac{\partial \log(\Gamma(\alpha))}{\partial \alpha} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$\beta_j = \frac{1}{\alpha_j} \frac{\sum_{i=1}^{n} x_i p(Z_i = j \mid x_i, \theta^m)}{\sum_{i=1}^{n} p(Z_i = j \mid x_i, \theta^m)} \quad (13)$$

3.3 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a criterion based on the concept of entropy used to calculate the goodness of fit of a model [12, 13]. The AIC calculation formula of GMM is as follows:

$$AIC = -2\log(L(\theta^*)) \quad (14)$$

$L(\theta^*)$ is log-likelihood function. The best model of Gaussian mixture is selected by the smallest AIC. The component number k of the best model is the classification class of the cluster analysis. The initial center is selected according to the range uniformity in the GMM model component.

4. COLOR CLUSTERING

4.1 Traditional K-Means Cluster Analysis

Clustering analysis is an unsupervised pattern recognition method. The data is classified based on the principle that "objects are clustered", and the same type of data objects have higher similarity [14]. The degree of similarity is usually expressed using a similarity coefficient or distance between each object.

K-means cluster analysis is one of the most commonly used cluster analysis methods. The algorithm has the advantages of being fast, intuitive and easy to implement, but it is necessary to determine the number of clusters in advance and to select the initial cluster center [15, 16]. The selected initial cluster center has a very large impact on the final classification results [16].

The steps of the k-means clustering analysis algorithm are as follows:

(1) Determining the number k of the classification of the data, and define an initial center for each class;

(2) Calculating the Euclidean distance between the data and the initial center as the similarity parameter, and classify the data by similarity parameter;

(3) Recalculating k centers according to the clustering results as various new centers;

(4) When k new centers are obtained, the Euclidean distance between each data and the initial center is recalculated as the similarity parameter, and iteratively loops until the clustering result is satisfied the indicator function.

4.2 Improved k-means clustering color segmentation algorithm

When applying the k-means clustering algorithm to color segmentation, it is necessary to determine the number of color image classification and the cluster initial center in advance. In order to avoid the problem that the color number k and the initial center are not properly selected and the clustering effect is not obvious or even wrong, a clustering algorithm based on multi-color space is proposed. After using the Gaussian mixture model to solve the clustering number and initial center, the simulated annealing optimization algorithm is used to improve k-means algorithm, which avoids the problem that the traditional k-means algorithm will fall into the local optimal solution. The core idea of the simulated annealing algorithm: In the process of searching, if the new solution is better than the currently searched optimal solution, the new solution is unconditionally accepted as the current optimal solution; if the new solution is worse than the current optimal solution, Then it is accepted with a certain probability, so the search process can jump out of the local optimal solution, so that the global optimal solution can be obtained[17]. The basic process is as follows:

(1) Initialization, giving the initial temperature and the initial solution, the target function value is calculated for the initial solution;

(2) Perturbing the initial solution to generate a new solution, and calculate its target value function again;

(3) If the objective function value of the new solution is less than or equal to the objective function value of the initial solution, accept the new solution and overwrite the current solution;

(4) Otherwise, accept the new solution with a certain probability;

(5) Cool down the current temperature and return to step (2) until the termination condition is met.

The specific steps of color image segmentation using the improved k-means clustering algorithm are as follows:

(1) Converting the image from RGB to HIS and CIELAB using equations (1), (2), and (3);

(2) Extracting the H component in the HIS and performing histogram data calculation;

(3) Gaussian mixture model is applied to the histogram data of H component, and solve it by EM algorithm. The optimal model is selected according to the AIC criterion to determine the cluster number k and the initial center.

(4) Extracting (a, b) two-dimensional components in the CIELAB space, and mapping the initial center which obtained in step (3) to (a, b);

(5) Dividing the data by class number k of the step (3) and the initial center in the step (4);

(6) The divided data is initialized to the simulated annealing algorithm, and the simulated annealing algorithm is used to obtain the clustering center. That is the global optimal solution;

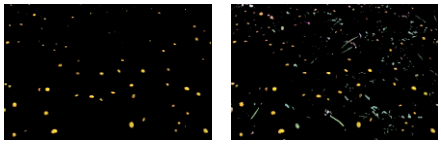
(7) Performing k-means cluster analysis using the clustering class number k in step (3) and the global optimal clustering center in step (6);

(8) Image color segmentation using clustering results.

5. EXPERIMENTAL RESULT

The clustering color segmentation experiment of color images is performed in the MATLAB environment of Windows system, and split the yellow color from the flower picture. The RGB image of the flower cluster is collected by the camera. As shown in Fig. 2.

Figure 2. Origin image of flowers.

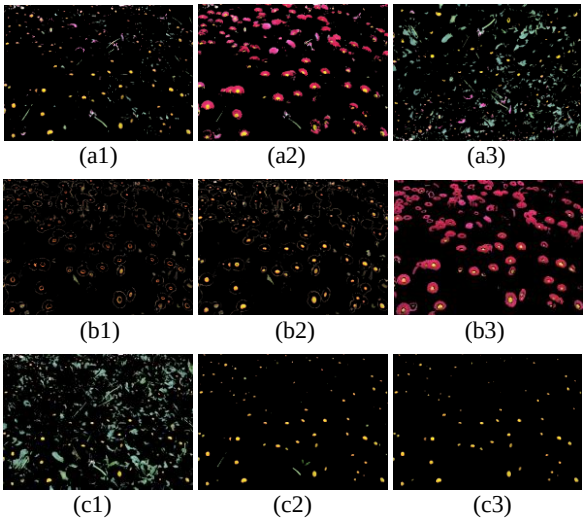


(1) (2)

Figure 3. The result of color segmentation by the proposed algorithm and traditional k-means algorithm

(1) The result of the proposed algorithm;(2) The result of the traditional k-means algorithm.

And by comparing the traditional k-means algorithm, as shown in Figure 3, it can be seen that the reslut of traditional k-means algorithm is mixed with a small amount of other colors. The improved k-means algorithm is better than the traditional k-means algorithm. In the RGB, HIS, LAB space, the algorithm is used to color-divide in the two components of the flower image, as shown in Figure 4. It can be seen from Figure 4 that the result of color segmentation in (a, b) components is best, which verifies the validity and rationality of the proposed algorithm. Then, the original image is separately subjected to image enhancement, brightness reduction, and brightness enhancement processing, and the three processed images are color-divided. The segmentation rate was calculated by calculating the number of yellow centers in the image. The number of yellow centers is shown in Table 1. It can be seen that the proposed algorithm effectively improves the robustness of segmentation.



(a1) (a2) (a3)

(b1) (b2) (b3)

(c1) (c2) (c3)

Figure 4. The result of color segmentation in RGB, HIS, Lab color space.

(a1) The result in (r, g); (a2) The result in (r, b); (a3) The result in (g, b); (b1) The result in (h, i); (b2) The result in (h, i); (b3) The result in (i, s); (c1) The result in (l, a); (c2) The result in (l, b); (c3) The result in (a, b);

Table 1. The rate of color segmentation

State of image		num	Rate (%)	Time (s)
Before the segmentation		85	--	--
After the segmentation	Original	76	89.41%	3.59
	Intensification	70	82.35%	3.89
	Brightness decrease	81	95.29%	3.78
	Brightness increased	82	96.47%	3.70

6. CONCLUSION

This paper proposes a k-means color segmentation algorithm for multi-color space, which effectively solves the parameters of classification and initial center need to be entered by using k-means clustering algorithm to segment colors. The H component in the HSI color space is used to determine the number of colors in the image. The number and the cluster center are determined by the Gaussian mixture model, and then the image is clustered and color segmented. Then, the simulated annealing algorithm is used to improve the traditional k-means algorithm, which avoids the problem that the traditional k-means algorithm is easy to fall into the local optimum. The experimental results show that the algorithm has good effect of color segmentation without human intervention and has good robustness.

7. REFERENCES

[1] Amruta B Patil. OTSU Thresholding Method for Flower Image Segmentation [J]. International Journal of Computational Engineering Research, 2016,06(05): 220 - 226.

[2] Guangzhu Xu, Xinyu Li, Bangjun Lei, et al. Unsupervised color image segmentation with color-alone feature using region growing pulse coupled neural network [J]. Neurocomputing,2018,306(6): 1-16.

[3] Sulaiman S N, Isa N A M. Adaptive fuzzy-K-means clustering algorithm for image segmentation[J]. Consumer Electronics IEEE Transactions on, 2010, 56(4):2661-2668.

[4] Haiyang Li, Hongzhou He, Yongge Wen. Dynamic particle swarm optimization and K-means clustering algorithm for image segmentation [J]. Optik,2015,126(24): 4817-4822.

[5] Hengyi Ren, Song He, Wenliang Chen. An improved k-means clustering algorithm is applied in image segmentation [J]. Communication technology, 2017,50(12): 2704-2707.

[6] Qian Wu, Minhui Wang. The measurement method of white spot area based on k-mean clustering of mixed color space [J]. New industrialization,2018,08(07): 88-93.

[7] Yurtsever, Ulas; Evirgen, Hayrettin; Avunduk, Mustafa Cihat. A New Augmented K-Means Algorithm for Seed Segmentation in Microscopic Images of the Colon Cancer[J]. Tehnicki Vjesnik. 2018,25(2): 382-389.

[8] Chunxi Cheng, Yanming Quan. Color image background subtraction based on HSI model [J]. Computer application,2009,29: 231-232+235.

[9] Huizhen Zhang, Yunyang Yan, Yi'an Liu, et al. Video flame detection based on superpixel segmentation and scintillation feature discrimination [J]. Data collection and processing, 2008,33(03): 512-520.

[10] Suyang Hu, Shuiyuan Huang, Zhiyi Chen. Digital photo background color replacement based on CIELAB color model [J]. Computer applications and software,2016,33(07): 229-233.

[11] Zhiyong Tao, Xiaofang Liu, Hezhang Wang. Gaussian mixture model clustering algorithm for fusion density peak [J]. Computer application,2018:1-7

[12] Jianan Wang. A kind of k-means clustering analysis based on AIC criteria [J]. Scientific outlook,2016,26(01): 198.)

[13] Tan M Y J, Biswas Rahul. The reliability of the Akaike information criterion method in cosmological model selection[J]. Monthly Notices of Royal Astronomical Society.2012,419(4): 3292-3303.

[14] Jun Wang, Shitong Wang, Zhaohong Deng. Several problems in cluster analysis [J]. Control and decision-making,2012,27(03): 321-328.

[15] Minchen Zhu, Weizhi Wang, Jingshan Huang. Improved initial cluster center selection in K-means clustering[J]. Engineering Computations.2014,31(8): 1661-1667.

[16] Jintao Li, Ping Ai, Zhaoxin Yue, et al. Improvement of k-means clustering algorithm [J]. Overseas electronic measurement technology, 2007,36(06): 9-13+21.

[17] Liguo Yao, Haisong Huang. Fault Diagnosis of Rolling Bearings Based on Improved K-Means Simulated Annealing Clustering Algorithm[J]. Modular Machine Tools & Automatic Processing Technology,2017(04):114-117.

A Hybrid DCT-CLAHE Approach for Brightness Enhancement of Uneven-illumination Underwater Images

Meng Ge
Department of Electrical and Electronic Engineering
Kyushu Institute of Technology
Fukuoka, Japan
mangorl1990@gmail.com

Qingqing Hong
Department of Electrical and Electronic Engineering
Kyushu Institute of Technology
Fukuoka, Japan
hongqingqing1221@gmail.com

Lifeng Zhang
Department of Electrical and Electronic Engineering
Kyushu Institute of Technology
Fukuoka, Japan
zhang@elcs.kyutech.ac.jp

ABSTRACT
Underwater images have been drawing sufficient attentions, as more and more researchers and companies are committed to ocean engineering, underwater archaeology, mining etc in recent years. Underwater images often suffer from unbalanced illumination, which bring image analysis and processing challenges. This paper introduces a novel approach for underwater image enhancement based on the DCT coefficient and CLAHE algorithm. Image changes with operations on the statistical model of the image DCT coefficients. The proposed methods firstly perform brightness equalization of the image, which optimize the underexposed and overexposed regions by regulating the low-frequency part of the image DCT coefficients. Afterwards CLAHE is applied to achieve image contrast enhancement. Finally experiments shows the better image enhancement performance compared to other algorithms.

CCS Concepts
• **Computer methodologies→Appearane and texture representations.**

Keywords
Underwater Image; Brightness Equalization; Image Enhancement; DCT Coefficient; CLAHE.

1. INTRODUCTION
In recent years, marine resource and ocean ecology have become topical issues with the rapid development of science and technology. Along with the rocketing demands for deep-sea research, underwater image process and enhancement are becoming more and more important, especially in the field of ocean engineering, monitoring sea life, aquatic species counting, underwater archaeology, robotics and mining [1][2]. It is not an easy work to capture clear underwater images due to the factor of underwater environment, such as light attenuation and scattering. Consequently underwater images often appear to be suffering from contrast degradation, color distortion and low visibility.

Since sunlight drops off within a certain distance and blue-green color dominates underwater images, artificial lighting are employed for capturing images [3]. If the image has been captured in an abnormal lighting condition, it leads to lower image quality [4]. The uneven illumination in deep sea brings image analysis challenges. To address these problems, image enhancement methods are imperative.

Image enhancement is called as the process of transforming the degraded content in the original image to better understandable image. Image enhancement process uses different methods to get the clarity in images which are often influenced by noise, blur and artifacts [4]. And the methods can be divided into two classes that perform process in spatial domain and frequency domain. The spatial domain refers to the image plane itself, and means in this category are based on direct manipulation of pixels in an image [5]. Histogram equalization (HE) is a very common method for image enhancement in spatial domain due to its simplicity and effectiveness [6]. As to the latter, the image is firstly changed to transformation domain and then inverse transformed to spatial domain after processing in the transformation domain. Discrete cosine transform (DCT) is usually applied in transformation domain, especially for image compression [7]. As directly applying HE methods to underwater images tends to lose details in dark regions and induce oversaturation in bright regions, a new combined method utilizing DCT coefficients and CLAHE algorithm is proposed to enhance the uneven-illumination underwater images in this paper.

2. Proposed Methodology
Noticeable differences often exist between the observed scene and the captured image, especially in deep sea. And the captured image may appear to be too darkened (under-exposure) or overly brightened (over-exposure) [8]. The method presented consists of two main steps. The first step is image brightness equalization by the means of DCT technique. And the second is image contrast enhancement utilizing CLAHE algorithm. The process of proposed method is shown in Figure 1.

2.1 YCbCr Color Space
The YCbCr color space is widely used for digital video and photography system. While the RGB represents color as red, green and blue components, the YCbCr stands for brightness and color different signals. In YCbCr color space, Y is luminance that occurs using black and white gray shades, and chrominance Cb and Cr are the blue-difference and red-difference components [9]. Since luminance is so sensitive but not the same to the

ICVIP 2018, December 29-31, 2018, Hong Kong, Hong Kong.

ACM ISBN 978-1-4503-6613-7/18/12…$15.00 .
http://doi.org/10.1145/3301506.3301539

chrominance, the YCbCr color space method makes use of its effect to show the variation of luminance and chrominance by sorting out the modules of the specified images [10]. The transformation from RGB to YCbCr is as follows:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.257 & 0.564 & 0.098 \\ -0.148 & -0.368 & -0.071 \\ 0.439 & -0.291 & 0.439 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

After the conversion, Y channel is extracted to perform the following Discrete Cosine Transformation to balance brightness of the image.

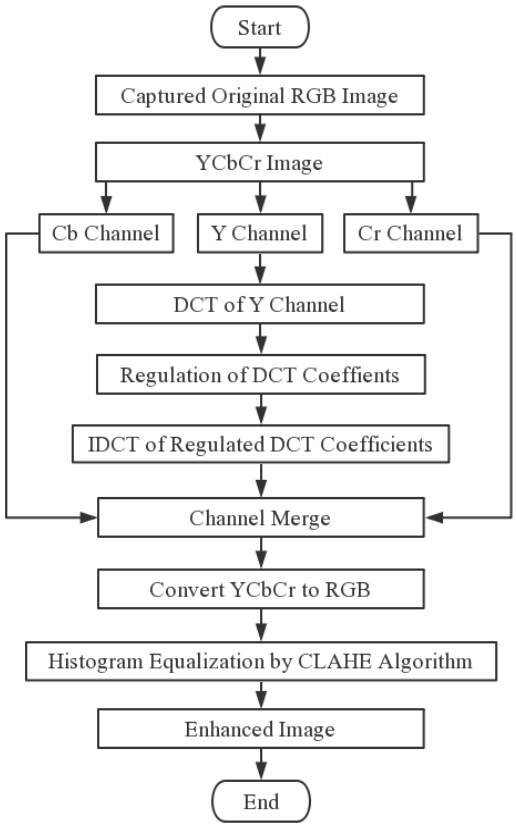


Figure 1. Flowchart of Proposed Method

2.2 Discrete Cosine Transform (DCT)

Discrete Cosine Transform (DCT) is usually applied in signal and image processing, especially for lossy compression on account of its strong energy compaction property [11][12]. DCT converts a signal from spatial domain to frequency domain. DCT is actual-valued and presents a superior estimate of an image with little coefficients [13]. The DCT and inverse DCT equation of an image is as formula (2) and formula (3):

$$F(i,j) = c(u)c(v)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\cos\left[\frac{(i+0.5)\pi}{N}u\right]\cos\left[\frac{(j+0.5)\pi}{N}v\right] \quad (2)$$

$$f(i,j) = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1}c(u)c(v)F(u,v)\cos\left[\frac{(i+0.5)\pi}{N}u\right]\cos\left[\frac{(j+0.5)\pi}{N}v\right] \quad (3)$$

Where:

$$c(u) = \begin{cases} \sqrt{\frac{1}{N}}, u = 0 \\ \sqrt{\frac{2}{N}}, u \neq 0 \end{cases} \quad (4)$$

An output DCT matrix is generated when the image is transformed by DCT. The larger values of coefficients are concentrated mostly in the upper left corner of the matrix (low-frequency), and values of coefficients tend to be smaller and smaller in the low right corner of the matrix (high-frequency). Since F (0, 0) as DC coefficients does not possess significant statistical regularity, this paper discusses statistical model of the AC coefficients. Fig 2 shows the Y channel of an YCbCr image and the DCT matrix of it.

The key point is that now high-frequency and low-frequency brightness information have been separated out by means of DCT. The proposed method achieves brightness equalization by regulating the low frequency coefficients to smaller values.

(a) **(b)**

Figure 2. (a) Y channel of an image (b) DCT matrix of a

2.3 Conventional Histogram Equalization (CHE)

Conventional histogram equalization (CHE) is a very popular algorithm for image enhancement owing to its simplicity and good performance. CHE tends to flatten and stretch the concentrated dynamic range of image's histogram to enhance the image with better contrast [14]. Suppose an image X has total pixels n and maximum gray level L. Then the probability of an occurrence of a pixel of level i in the image X is

$$p_x(i) = \frac{n_i}{n}, 0 \le i < L \quad (5)$$

And $p_x(i)$ is actually is the histogram for pixel value I, which normalized to [0,1].

The corresponding Cumulative Density Function (CDF) is defined as

$$CDF_x(i) = \sum_{j=0}^{i} p_x(j) \quad (6)$$

A transformation is created to produce a new image Y with a flatter histogram. And CDF of the new image would be linearized across the value range like

$$CDF_y(i) = iK \quad (7)$$

For some constant K. The output image of CHE can be expressed as

$$y = CDF_x(x) \qquad (8)$$

Where y is always normalized into the range [0, 1].

2.4 Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE differs from conventional histogram equalization in its contrast constraining. And CLAHE was developed to prevent the over amplification of noise that in CHE [15]. The algorithm consists of few steps, which including: get every one of the inputs; pre-process the inputs; prepare each logical locale to deliver gray level mappings; interpolate gray level mapping keeping in mind the end goal to merge final CLAHE picture [16]. The flowchart of CLAHE is presented in Figure 3.

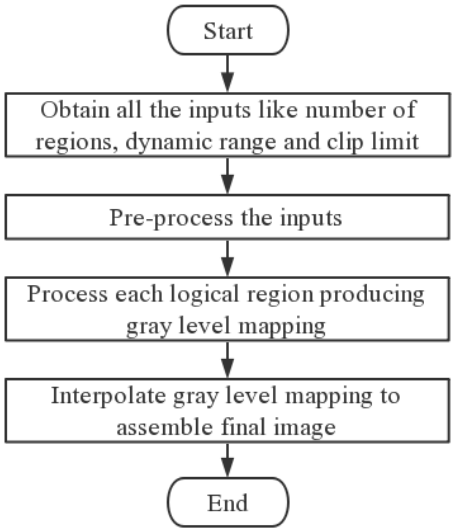


Figure 3. Flowchart of CLAHE

In consequence, CLAHE is combined into our approach for contrast enhancement of brightness-equalized underwater images. The hybrid DCT-CLAHE method is accordingly proposed. As can be seen from the following Figure 4, compared to the mere DCT technique, performance of the image get considerable improved through uniting CLAHE algorithm.

(a) **(b)**

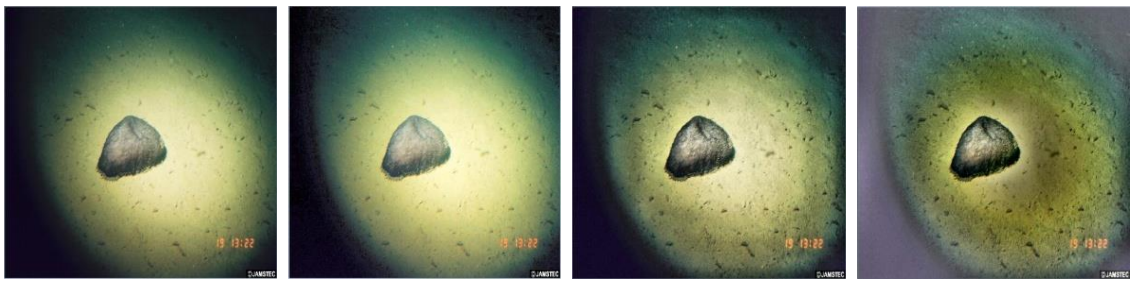
Figure 4. (a) image after brightness equalization by DCT technique (b) image processed by DCT-CLAHE technique

3. Results and Discussion

The experimental analysis is performed on underwater images. Representative uneven-illumination underwater color images are used for performance evaluation. Mean Luminance Error (MLS) is set as a quantitative criterion to estimate the performance of brightness equalization, which is the smaller, the better. Results of quantitative comparison are presented in Table 1, while the visual contrast is shown in Figure 5~8. It is quite clear that the proposed method performs much better in equalizing the brightness and preserving the contrast of underwater images simultaneously.

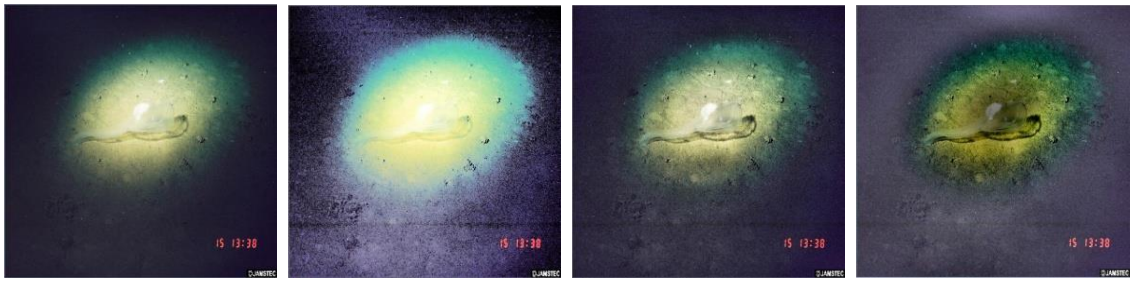
Table 1. Comparison of MLE

MLE	Original image	CHE	CLAHE	Proposed method
Example1	213.6	234.8	211.6	19.2
Example2	236.9	237.2	230.9	46.4
Example3	191.4	234.4	201.7	84.3
Example4	202.9	238.9	193.9	31.2



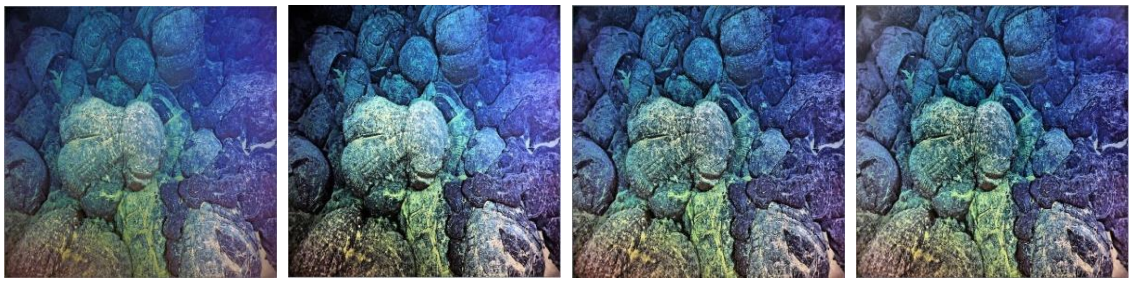
(a) Original image **(b) CHE** **(c) CLAHE** **(d) Proposed method**

Figure 5. Comparison of example image 1



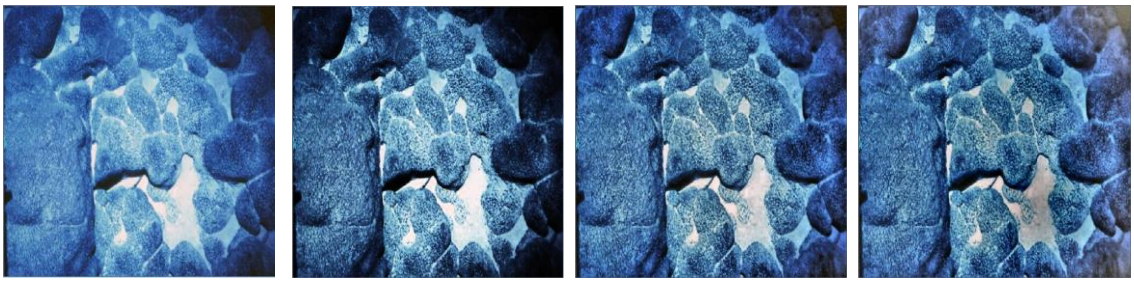
(a) Original image **(b) CHE** **(c) CLAHE** **(d) Proposed method**

Figure 6. Compasion of example image 2



(a) Original image **(b) CHE** **(c) CLAHE** **(d) Proposed method**

Figure 7. Comparison of example image 3



(a) Original image **(b) CHE** **(c) CLAHE** **(d) Proposed method**

Figure 8 Comparison of example image 4

4. CONCLUSION

The uneven illumination in deep sea brings underwater image process inconvenience and challenges. In this paper, a novel approach for optimizing underwater image brightness has been presented. Our method is used to enhance the too darkened or overly brightened regions of images by regulating the low-frequency part of image DCT coefficients. And in the meanwhile, image enhancement performs well in contrast on account of CLAHE technique. From the above experiments and results, the proposed approach is proved to be a more effective and convenient way for uneven-illumination underwater image enhancement.

5. REFERENCES

[1] Erickson Nascumento, Mario Campos, and Wagner Barros. 2009. Recovery of Underwater Scenes from Automatically Restored Images. In *XXII Brazilian Symposium on Computer Graphics and Image Processing*, 330-337.

[2] Timm Schoening, Daniel Langenkamper, Bjorn Steinbrink, Daniel Brun, and Tim W. Nattkemper. 2015. Rapid image processing and classification in underwater exploration using advanced high performance computing. In *OCEANS 2015-MTS/IEEE Washington* (Washington DC, USA, Oct, 2015), 1-5. DOI= http://doi.org/10.23919/OCEANS.2015.7401952

[3] Srividhya K, and Ramya M.M. 2015. Performance of analysis of pre-processing filters for underwater images. In *International Conference on Robotics, Automation, Control and Embedded Systems(RACE)*(Chennai, India, Feb, 2015), 1-7.DOI= http://doi.org/10.1109/RACE.2015.7097234

[4] Srividhya K, and Ramya M.M. 2015. Performance of analysis of pre-processing filters for underwater images. In *International Conference on Robotics, Automation, Control and Embedded Systems(RACE)*(Chennai, India, Feb, 2015), 1-7.DOI= http://doi.org/10.1109/RACE.2015.7097234

[5] Rafael C. Gonzalez. 1992. Digital Image Processing.

[6] Joung-Youn Kim, Lee-Sup Kim, and Seung-Ho Hwang. 2000. In 2000 IEEE International Symposium on Circuits and System(Geneva, Switzerland, May, 2000), 537-540. DOI= http://doi.org/10.1109/ISCAS.2000.858807

[7] Fatma Harman, and Yucel Kocygit. 2017. A new approach to genetic algorithm in image compression. In *10th International Conference on Electrical and Electronics Engineering* (Bursa, Turkey, Dec, 2017), 894-898.

[8] Chengho Hsin, Nien-Tsu Hung, Jui-Jung Kao, and Shaw-Jyh Shin. 2010. Improving image luminance appearance through recurrent local intensity adaptation. In *International Conference on Signal Processing Systems* (Dalian, China,

July, 2010), 31-35. DOI= http://doi,org/10.1109/ICSPS.2010.5555227

[9] Aniket Roy, Arpan Kumar Maiti, and Kuntal Ghosh. 2015. A perception based color image adaptive watermarking scheme in YCbCr space. In *International Conference on Signal Processing and Integrated Networks* (Noida, India, Feb, 2015), 537-543. DOI= http://doi.org/10.1109/SPIN.2015.7095399

[10] S.Gopinathan, and S.Gayathri. 2016. A Study on Image Enhancement Techniques using YCbCr Color Space Methods. In *International Journal of Advanced Engineering Research and Science(IJAERS)*, 105-112. DOI= http://doi.org/10.22161/ijaers.3.8.4

[11] M. Narasimha, and A. Peterson. 1978. On the Computation of the Discrete Cosine Transform. In *IEEE Transactions on Communications* (June, 1978), 934-936. DOI= http://doi.org/10.1109/TCOM.1978.1094144

[12] J. Makhoul. 1980. A fast cosine transform in one and two dimensions. In *IEEE Transactions on Acoustics, Speech, and Signal Processing* (Feb, 1980), 27-34. DOI= http://doi.org/10.1109/TASSP.1980.1163351

[13] Nisheeth Dubey, Sandeep Kumar Tiwari, and Pankaj Sharma. 2017. A hybrid DCT-DHE approach for enhancement of low contrast underwater images. In *International Conference on Recent Innovations in Signal processing and Embedded Systems* (Bhopal, India, Oct, 2017), 304-309. DOI= http://doi.org/10.1109/RISE.2017.8378171

[14] Chen Soong Der, and Manjit Singh Sidhu. 2009. Re-evaluation of Autoatic Global Histogram Equalization-based Contrast Enhancement Methods. In *Electronic Journal of Computer Science & Information Technology* (May, 2009), 13-17.

[15] Rajesh Garg, Bhawna Mittal, and Sheetal Garg. 2011. Histogram Equalization Techniques for Image Enhancement. In *International Journal of Electronics & Communication Technology* (March, 2011), 107-111.

[16] Ankit Chaubey, and Anshul Atre. 2017. A hybrid DWT-DCLAHE method for enhancement of low contrast underwater images. In International conference of Electronics, Communication and Aerospace Technology (Coimbatore, India, April, 2017), 196-201. DOI= http://doi.org/10.1109/ICECA.2017.8203670

Exemplar-Based Image Inpainting using Automatic Patch Optimization

Xuehui Bi[1], Huaming Liu[1,2], Guanming Lu[2],Jian Wei[2],Yan Chao[1]

Fuyang Normal University
No.100, Qinghe West Road
Fuyang,Anhui,China
236037(0865582591393)
bixuehui888@163.com

Nanjing University of Posts and Telecommunications
No.66, Xinmofan Road
Nanjing, Jiangsu, China
210003(0862583492615)
liuhuaming888@126.cm

ABSTRACT

Image inpainting has a wide range of applications in image processing. Exemplar-based technique can inpaint texture and structure regions simultaneously. However, the size of patch influences the result of inpainting. So far, there is no easy way to automatically determine the size of patch. Structure tensor can be used to determine priority, because it is able to determine the properties of a local area. In the paper, the size of patch is computed by structure tensor when adopting exemplar-based technique. To reduce blockiness, boundary constraints is added for searching for similar patches. The experiments show our proposed method can inpaint texture, flat, structure images. Adding boundary constraints can eliminate blockiness to a certain extent.

CCS Concepts

•Computing methodologies →Computer graphics → Image manipulation → Image processing.

Keywords

Boundary constraint; exemplar size; inpainting; structure tensor.

1. INTRODUCTION

Image restoration has been paid more and more attention in image editing, which has caused many scholars to study image restoration. The restoration of denoising and deblurring belong to blind repair. The restoration of filling damaged regions and removing objects are non-blind repair (or called inpainting). Bertalmio et al. [1] proposed image inpainting for repairing digital images which were damaged for some reasons such as scratches, creases, color shedding and so on. Then more and more scholars pay attention to digital image inpainting.

There are two typical repair methods for image inpainting. One is partial differential equation (PDF) inpainting[2-6], since partial differentials equation can be derived from the variation, the variational method also belongs to PDF method[7]. The other is exemplar-based inpainting[8-11]. When using PDF-based regularization, the missing regions are filled by diffusion through the surrounding information. Since PDF is only suitable for repairing small areas, blurring may occur when the missing area is large. Exemplar-based technique can inpaint large damaged regions. Since it searches for best similar patches to fill missing regions and preferentially fill the structure regions, in many cases, satisfactory results can be obtained. The order of filling is very important, and if the image has noise or irregular texture, it may lead to instability of priorities calculation, so that satisfactory repair results can not be obtained. Criminisi et al. [8] proposed priority using gradient-based data term, it only use a pixel in the center of the patch, data term is susceptible to noise, which will make the priority calculation inaccurate and affect the fill order. Meur et al.[10] proposed a novel priority calculation using structure tensors $\nabla I_i \nabla I_i^T$, which consider edge regions, corner regions and flat regions. Texture regions are not fully considered, the calculation of priority will be inaccurate and fail to inpaint image. Xu et al.[11] compute priority using sparsity-based structure term, which consider local neighbor region of target patch. The structure sparsity of patch have different value on corner, edge, texture and flat regions, which helps to characterize the nature of the target patch. However, irregular textures may be mistaken for boundaries, and it will spend most time to compute distance with neighboring patches.

In [8] [10] [11], the priority denotes the property of the target patch, if the value of priority is higher, there may be an edge, which should be filled firstly. In fact, the properties of the patch can also determine the size of the block. Since the size of patch will also affect the quality of the repair [12-14], in [13; 14], they use the size of patch from 3×3 to 15×15 for optimizing inpainting. However, they will spend more time to choose the size of the patch.

In fact, the properties of the patch can also determine the size of the patch in our experiments. In the paper, structure tensor $\nabla I_i \nabla I_i^T$ is used to determine the size of patch. The flat region has low structure tensor value, more large size should be used to fill the target patch. In the opposite case, the larger value of the structure tensor, the smaller size of the patch is used to fill. Sparsity-based structure [11] will spend more time for computation than structure tensor, so structure tensor can take less time to fill missing regions.

The essay has been organized in the following way. Section 2 presents related work. Automatic patch optimization is introduced in section 3. The results of experiment presents in section 4. The main work is concluded in the section 5.

2. RELATED WORK

2.1 Exemplar-Based Technique

Our work is based on exemplar technique which is proposed by Criminisi. The process of inpainting is shown in Fig.1. Firstly, find

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301549

the edge of filling $\partial\Omega$ as shown in Fig.1(a). Secondly, compute priority of patch to be filled (target patch) as shown in Fig.1(b), the priority can be calculated by (1).

$$P(p)=C(p)D(p) \tag{1}$$

Where $P(p)$ is priority of patch (ψ_p) whose center is pixel p . $C(p)$ and $D(p)$ are confident term and data term, respectively. They are computed by (2) and (3), respectively.

$$C(p)=\frac{\sum_{q\in\hat{\psi}_p} C^0(q)}{|\psi_p|}, \hat{\psi}_p=\psi_p\cap S \tag{2}$$

Where S denotes the known pixels in the image. $\hat{\psi}_p$ represents the known pixels in the ψ_p . The term $|\psi_p|$ refers to the number of pixel in the ψ_p . $C^0(q)$ is initialized value in the image, when $\forall q\in S$, $C^0(q)$=1 , and $\forall q\in\Omega$, $C^0(q)$=0 . I denotes image set, $I=S\cup\Omega$.

$$D(p)=\frac{|\nabla I_p^{\perp}\cdot n_p|}{\alpha} \tag{3}$$

α refers to a normalization factor(α=255 for gray image), n_p is a unit vector orthogonal to the edge $\partial\Omega$ in the center p . $\perp$ stands for orthogonal operator.

Thirdly, search for similar patch ($\psi_{q'}, \psi_{q''}, \cdots$) in the sample region (S) as shown in Fig.1(c). The similar patch can be found by distance of sum of squared differences (SSD) between two patches by (4).

$$D_{ssd}(\psi_p,\psi_q)=\frac{1}{N}\sum_{c=1}^{m}\sum_{\substack{i=1\\ \psi_{pi}\in\psi_p\cap S\\ \psi_{qi}\in S}}^{N}(\psi_{pi}-\psi_{qi})^2 \tag{4}$$

Where the term m=3 has been used to refer to situations in color image. N denotes the known pixels in the ψ_p .The patch which has the smallest SSD is the best similar patch. Formally,

$$\psi_{\hat{q}}=\arg\min_{\psi_q\in S} D_{ssd}(\psi_p,\psi_q) \tag{5}$$

Fourthly, fill the target patch using best similar patch as shown in Fig.1 (d). Then the confidence term of filled boundary is updated after filling target patch.

2.2 Structure tensor

The boundary in the image can be presented by structure tensor as introduced in [15]. The structure tensor can reflect the local geometry information. In [10], structure tensor was used to define the data term for calculating priority. Its definition is obtained by (6).

$$J_\rho=K_\rho * \sum_{c=1}^{3}\nabla I_c \nabla I_c^T \tag{6}$$

Where ∇I_c is the gradient of the c th color channel, K_ρ refers to the Gaussian kernel whose standard deviation is ρ . T is the transpose. J_ρ is positive semi-definite, the eigenvalues of structure tensor J_ρ can be computed by:

$$\lambda_{1,2}=\frac{1}{2}[J_{11}+J_{22}\pm\sqrt{(J_{11}-J_{22})^2+4J_{12}^2}] \tag{7}$$

The corresponding orthogonal eigenvectors can be calculated by:

$$v_1=\begin{bmatrix}2J_{12}\\ J_{22}-J_{11}+\sqrt{(J_{11}-J_{22})^2+4J_{12}^2}\end{bmatrix}, v_1\perp v_2 \tag{8}$$

The larger eigenvalue represents the strength of the local image boundary, the corresponding eigenvector points across the boundary. In the flat region, λ_1 and λ_2 are smaller and $\lambda_1\simeq\lambda_2$; in edge region, λ_1 is larger, λ_2 is smaller and $\lambda_1\gg\lambda_2$;in corner region, λ_1 and λ_2 are larger and $\lambda_1\simeq\lambda_2$. Fig.2 shows the v_1 and v_2 .

3. PROPOSED APPROACH

3.1 Formulation and motivation

In this paper, exemplar-based inpainting technique is extended, because it can inpaint texture and structure simultaneously and fill large missing regions. The size of patch, however, will influence inpainting effect. Here patches of different sizes are chosen to optimize results from 3×3 to 15×15 in [13; 14], while, unfortunately, the process of inpainting will spent more time in searching similar patches. Can the size of patches be automatically selected? Inspired by structure tensor, in the flat region, large size should be chosen, and in the edge and corner, small size should be

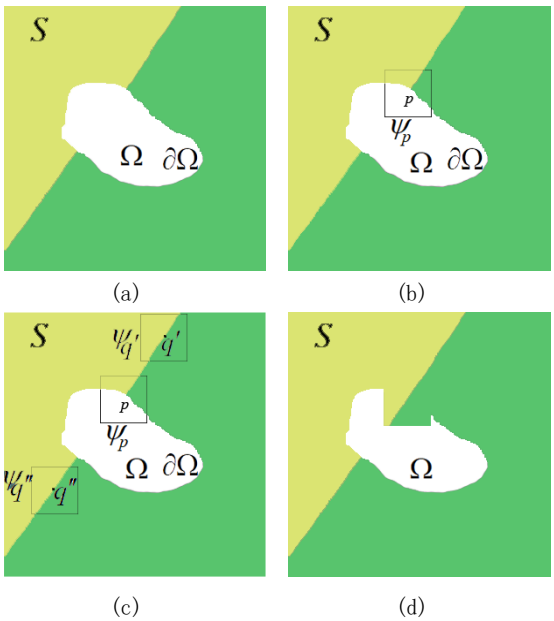


Figure 1. Exemplar-based technique. (a) Target region Ω ; (b) Choose a patch to be filled; (c) Search for similar patches; (d) Fill the target patch.

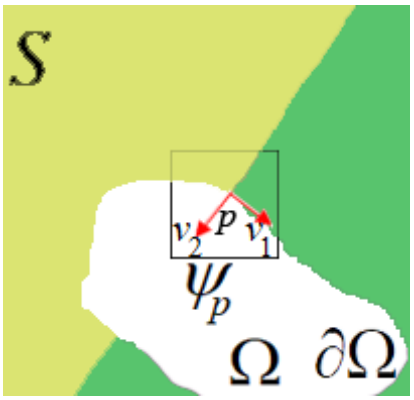


Figure 2. Structure tensor of point p

chosen. The size of patch should be determined by λ_1 and λ_2. In [10], the data term is defined as follows:

$$S(p)=\alpha+(1-\alpha)\exp\left(-\frac{\eta}{(\lambda_1-\lambda_2)^2}\right) \quad (9)$$

Where the term $\alpha\in 0,1$ and η is a positive value. $S(p)\in 0,1$, the larger the value of $S(p)$ is, the stronger the boundary of the local region is.

3.2 Automatic patch size selection

To better select the size of the patch, (9) is modified as:

$$S(p)=\alpha+(1-\alpha)\exp\left(-\frac{\eta}{|\lambda_1-\lambda_2|}\right) \quad (10)$$

the size of patch is set from 7×7 to 25×25. Assume α=0.3 and η=300. Polynomial is used to fit the size of patches. The computation of size of patch is defined by (11).

$$Size(p)=a_4S(p)^4+a_3s(p)^3+a_2s(p)^2+a_1s(p)^1+a_0 \quad (11)$$

Where term $Size(p)$ is the size of patch. Assume a_4,a_3,a_2,a_1,a_0 = -127.3716,349.1306,-349.9226,135.7657,-0.6111.

Table 1 shows the size of patch $S(p)$ when $S(p)$ varies from 0.3 to 1. $S(p)$ needs to be post processed sometimes, such as four let five enter (round($Size(p)$)) and be taken odd number operation $Odd(Size(p))$. Fig.3 shows the calculation curve of patch size. The larger $S(p)$ is, the smaller the size of the block is, which fits our block selection strategy.

3.3 Boundary constrain similar patch

Boundary matching distance (BMD) was employed in the video process [16; 17]. The target patch is filled by similar patch. The boundary ($B_{p_s}^{out}$ as shown in Fig.4(c)) should be close to its filling boundary ($B_{q_t}^{in}$ as shown in Fig.4(c)). Fig.4 shows the process of boundary constrain. ψ_p is the target patch, the candidate patch ψ_q is its similar patch as shown in Fig.4(a). ψ_{q_t} is used to fill ψ_{p_t}. The boundary distance (BMD) is defined by(12).

$$D_{bmd}(\psi_p,\psi_q)=\frac{1}{M}\sqrt{\sum_{c=1}^{3}\sum_{m=1}^{M}(B_{ps}^{out}(m)-B_{qt}^{in}(m))^2} \quad (12).$$

Where M is number of boundary points. When $c=3$, the image denotes color image. $B_{q_t}^{in}$ is edge filled in the ψ_{p_t}, $B_{p_s}^{out}$ is the outer boundary of the ψ_{p_t}. When finding similar patches, distance of patches is defined by (13).

$$D(\psi_p,\psi_q)=\omega D_{ssd}+(1\text{-}\omega)\ D_{bmd} \quad (13)$$

Where D_{ssd} is computed by (4). ω is the weight. Fig.5 gives the signal-to-noise ratio (PSNR) value for testing 11 images. When $\omega=0.45$, the average value of PSNR reaches the maximum. Assume $\omega=0.45$ as the default value.

Table 1. The computation of patch size

$S(p)$	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
$Size(p)$	17.0	16.8	15.5	13.8	12.1	10.6	9.0	7
round($Size(p)$)	17	17	16	14	12	11	10	7
Odd($Size(p)$)	17	17	17	15	13	12	11	7

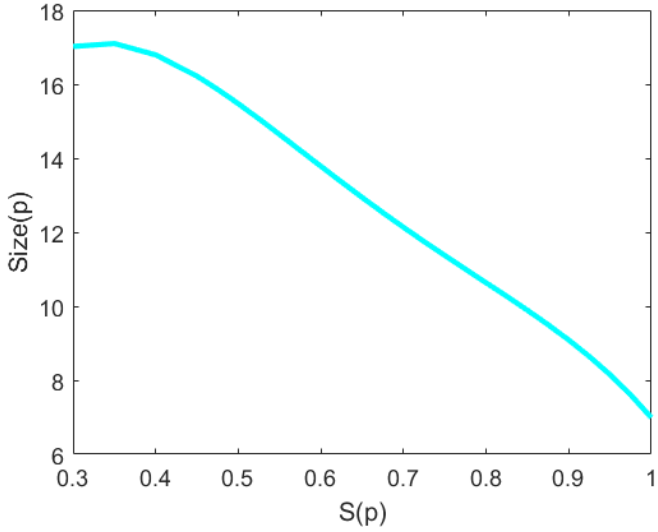


Figure 3. The calculation curve of patch size.

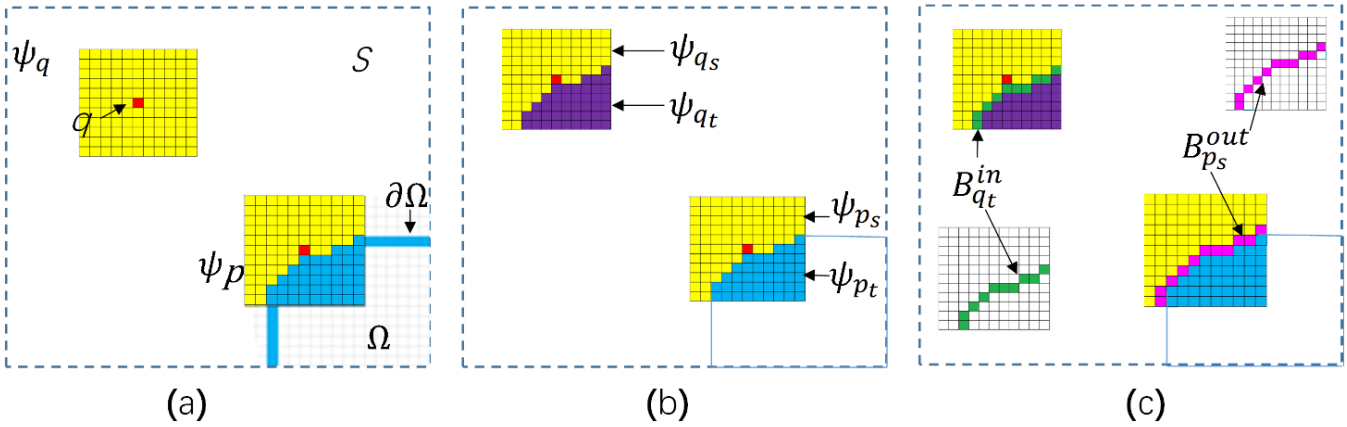


Figure 4. The boundary constrain of inpainting. (a) ψ_q is a candidate patch of ψ_p. (b) target patch ψ_{p_t} is filled by ψ_{q_t}. (c) $B_{q_t}^{in}$ is edge filled in the ψ_{p_t}, $B_{p_s}^{out}$ is the outer boundary of the ψ_{p_t}.

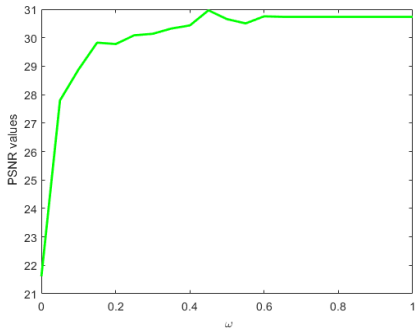


Figure 5. The value of PSNR varies from different ω.

4. RESULTS

Fig6 – Fig.8 show the results of inpainting by our method and other methods [8; 12; 18-22]. Fig.6(c)-6(j) and Fig.7(c)-7(j) are the results of inpainting by Criminisi's method[8], Deng's method[18], Anupam's method[19], Jurio's method[20], Barnes's method[21], Newson's method[12], Fedorov's method[22], and our method. In Fig.6, the target region contains structure and texture, especially irregular textures such as water. So there are some difficulties in repairing. Fig.6(c), Fig.6(d) and Fig.6(f) show the bad results, Fig.6(e) , Fig.6(g)-6(j) show the pleasing results. Multiresolution inpainting is used in both Fig.6(g) and Fig.6(h). However, our method only use single image and obtain plausible result.

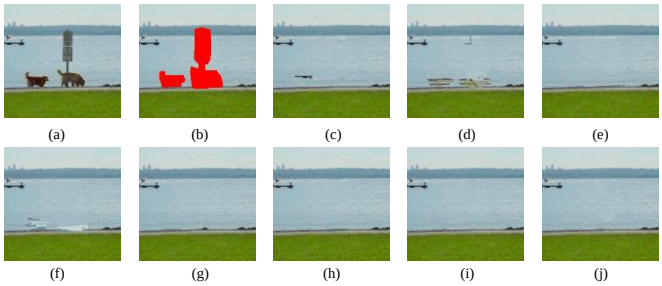


Figure 6. Object removing. (a) Origin image; (b)Mask image; (c)Criminisi's method[8];(d) Deng's method[18];(e) Anupam's method [19]; (f) Jurio's method[20]; (g)Barnes's method[21]; (h)Newson's method[12];(i) Fedorov's[22]; (j)Our method.

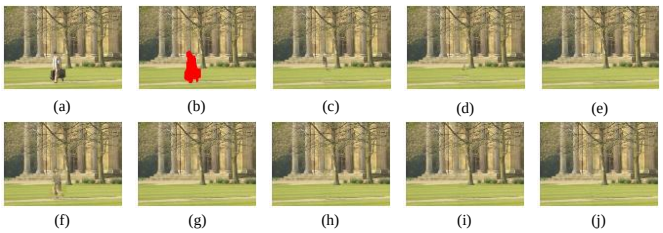


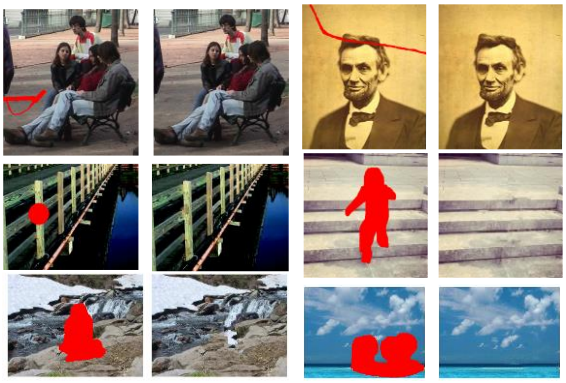
Figure 7. Object removing. (a) Origin image; (b)Mask image; (c)Criminisi's method[8];(d) Deng's method[18];(e) Anupam's method [19]; (f) Jurio's method[20]; (g)Barnes's method[21]; (h)Newson's method[12];(i) Fedorov's[22]; (j)Our method.

Fig.7 contain irregular texture regions such as grass, and structure regions such as road. Repairing is more difficult. Fig.7 show the same plausible results by [12; 19; 21; 22] and our method. Fig.7(c), Fig.7(d) and Fig.7(f) show the unreasonable result by Criminisi's method[8], Deng's method[18] and Jurio's method[20], respectively. This shows that the proper size of the block can improve the repair effect.

Fig.8 show more examples of repairing. Our algorithm is capable of repairing different types of images, including flat, structural and textured images as shown top row, middle row and bottom row respectively. The results of inpainting are visually pleasing results. Fig.9 show the results with and without boundary constrain. Fig.9(c) shows the more pleasing result than Fig.9 (b). This shows that boundary constraint can improve the repair effect and eliminate some blockiness, but the blockiness can not be completely removed as shown in Fig.9(c).

5. CONCLUSION

A novel method is proposed, which can choose the size of patch automatically when using exemplar-based technique. Since the structure tensor can judge the property of local region, it can be used to guide the selection of patch size. Experimental results demonstrate the effectiveness of our method. To reduce blockiness, boundary constrain is added for searching similar patches, the comparison of with and without boundary constraint show that boundary constraint can further reduce blockiness. In addition, boundary constraint should use different weights for textures, flat, and structural regions. Our next step work is to automatically determine the weight and further improve the repair effect.



(a) Mask (b) Proposed (a) Mask (b) Proposed

Figure 8. The inpainting results by proposed method.

(a) (b) (c)

Figure 9. Our method's results with and without boundary constrain. (a) Mask; (b) without boundary constrain; (c) with boundary constrain.

6. ACKNOWLEDGMENTS

This research was funded by the Key projects of natural science research in Anhui colleges and Universities(No.KJ2018A0345), the key fund projects of Young Talents in Fuyang Normal University (No.rcxm201706), the National Natural Science Foundation of China(61772430), Postgraduate Research & Practice Innovation Program of Jiangsu Province(No.KYCX18_0901), the Natural Fund Project in Fuyang Normal University(No.2017FSKJ17). This work is also supported by the Key Research and Development Program of Jiangsu Province (No.BE2016775). The 2017 horizontal cooperation project of Fuyang municipal government-Fuyang Normal College (No. XDHX201732).

7. REFERENCES

[1] Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C., 2000. Image inpainting. *In Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (New Orleans, LA, July 23-28, 2000)ACM, New York, NY, 417-424. DOI= http://doi.acm.org/10.1145/344779.344972.

[2] Chan, T.F., 2001. Nontexture inpainting by curvature-driven diffusions. *J. Visual Commun. Image Represent*. 12, 4 (Dec. 2001), 436-449. DOI= https://doi.org/10.1006/jvci.2001.0487.

[3] Huang, F., Chen, Y.M., Duensing, G.R., Akao, J., Rubin, A., and Saylor, C., 2005. Application of partial differential equation-based inpainting on sensitivity maps. *Magn. Reson. Med*. 53, 2 (Feb. 2005), 388-397. DOI= https://doi.org/10.1002/mrm.20346.

[4] Li, P., Li, S.J., Yao, Z.A., and Zhang, Z.J., 2013. Two anisotropic fourth-order partial differential equations for image inpainting. *IET Image Processing*. 7, 3 (Apr. 2013), 260-269. DOI= https://doi.org/10.1049/iet-ipr.2012.0592.

[5] Li, S.J. and Yao, Z.A., 2013. Image inpainting algorithm based on partial differential equation technique. *Imaging Sci. J.* 61, 3 (Mar. 2013), 292-300. DOI= https://doi.org/10.1179/1743131x11y.0000000055.

[6] Li, S.J. and Yang, X.H., 2017. Novel image inpainting algorithm based on adaptive fourth-order partial differential equation. *IET Image Processing* 11, 10 (Oct. 2017), 870-879. DOI= https://doi.org/10.1049/iet-ipr.2016.0898.

[7] Aubert, G. and Kornprobst, P. 2002. Mathematical Problems in Image Processing.Partial Differential Equations and the Calculus of Variations. *Applied Mathematical Sciences*.147, Springer, New York.

[8] Criminisi, A., Pérez, P., and Toyama, K., 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process*. 13, 9 (Sept. 2004), 1200-1212. DOI= http://doi.acm.org/10.1109/TIP.2004.833105.

[9] Shih, T.K., Tang, N.C., and Hwang, J.N., 2009. Exemplar-Based Video Inpainting Without Ghost Shadow Artifacts by Maintaining Temporal Continuity. *IEEE Trans. Circuits Syst. Video Technol.* 19, 3 (Mar. 2009), 347-360. DOI= http://doi.acm.org/10.1109/TCSVT.2009.2013519.

[10] Le meur, O., Gautier, J., and Guillemot, C., 2011. Examplar-based inpainting based on local geometry. *In Proceeding of the 18th IEEE International Conference on Image Processing* (Brussels, Belgium, Sep. 11-14, 2011), IEEE, New York, NY, USA, 3401-3404. DOI= http://doi.acm.org/10.1109/ICIP.2011.6116441.

[11] Xu, Z. and Sun, J., 2010. Image inpainting by patch propagation using patch sparsity. *IEEE Trans. Image Process* 19, 5, (May. 2010), 1153-1165.DOI= http://doi.acm.org/10.1109/TIP.2010.2042098.

[12] Newson, A., Almansa, A., Gousseau, Y., and Pérez, P., 2017. Non-local patch-based image inpainting. *Image Processing On Line* 7, ()373-385. DOI= https://doi.org/10.5201/ipol.2017.189.

[13] Borole, R.P. and Bonde, S.V., 2017. Image Restoration and Object Removal Using Prioritized Adaptive Patch-Based Inpainting in a Wavelet Domain. *J. Inf. Process. Syst.* 13, 5. 1183 ~ 1202. DOI=https://doi.org/10.3745/jips.02.0031.

[14] Borole, R.P. and Bonde, S.V., 2017. Image Restoration using Prioritized Exemplar Inpainting with Automatic Patch Optimization. *Journal of The Institution of Engineers (India): Series B* 98, 3(Jun. 2017), 311-319. DOI= https://doi.org/10.1007/s40031-016-0268-y.

[15] Di Zenzo, S., 1986. A note on the gradient of a multi-image. *Computer vision, graphics, and image processing* 33, 1, (Jan.

1986), 116-125. DOI=https:/doi.org/10.1016/0734-189x(86)90223-9.

[16] Wang, Y.-K., Hannuksela, M.M., Varsa, V., Hourunranta, A., and Gabbouj, M., 2002. The error concealment feature in the H. 26L test model. *In Proceeding of International Conference on Image Processing.* (Rochester, NY, USA, Sept.22-25, 2002) IEEE, New York, NY. II-II. DOI= https://doi.org/10.1109/icip.2002.1040054

[17] Wang, J., Tang, Y., Li, S., Ishiwata, S., and Goto, S., 2009. Side match distortion based adaptive error concealment order for 1Seg video broadcasting application. *In Proceeding of 2009 IEEE International Symposium on Circuits and Systems*, (Taipei, Taiwan, May. 24-27, 2009) IEEE,New York, NY, USA, 33-136.DOI= https://doi.org/10.1109/iscas.2009.5117703.

[18] Deng, L.J., Huang, T.Z., and Zhao, X.L., 2015. Exemplar-Based Image Inpainting Using a Modified Priority Definition. *Plos One* 10, 10, (Oct. 2015), 1-18. DOI= https://doi.org/10.1371/journal.pone.0141199.

[19] Anupam, Goyal, P., and Diwakar, S., 2010. Fast and enhanced algorithm for exemplar based image inpainting. *In Proceeding of 2010 Fourth Pacific-Rim Symposium on Image and Video Technology*, (Singapore, Singapore, Nov.14-17, 2010), IEEE, New York, NY, USA, 325-330. DOI= https://doi.org/10.1109/psivt.2010.61.

[20] Jurio, A., Paternain, D., Pagola, M., Marco-Detchart, C., and Bustince, H., 2017. Two-step Algorithm for Image Inpainting. *In Proceeding of Advances in Intelligent Systems and Computing*, (Warsaw, Poland, Sep. 11-15, 2017), Springer, Cham, Switzerland, 302-313. DOI=https://doi.org/10.1007/978-3-319-66824-6_27.

[21] Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D.B., 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* 28, 3(Aug. 2009), 24:1-24:11. DOI= https://doi.org/10.1145/1531326.1531330.

[22] Fedorov, V., Facciolo, G., and Arias, P., 2015. Variational framework for non-local inpainting. *Image Processing On Line* 5, () 362-386. DOI= https://doi.org/10.5201/ipol.2015.136

Analysis of Chromatic Characteristics, in Satellite Images for the Classification of Vegetation Covers and Deforested Areas

Wilver Auccahuasi
Universidad Continental
Huancayo, Perú
wauccahuasi@continental.edu.pe

Madelaine Bernardo
Instituto Peruano de Investigacion en Ingenieria Avanzada
Lima, Perú
mbernardo@bioingenieriaperu.edu.pe

Elizabeth Oré Núñez
Universidad Continental
Huancayo, Perú
eore@continental.edu.pe

Fernando Sernaque
Instituto Peruano de Investigacion en Ingenieria Avanzada
Lima, Perú
fsernaque@bioingenieriaperu.edu.pe

Percy Castro
Instituto Peruano de Investigacion en Ingenieria Avanzada
Lima, Perú
pcastro@bioingenieriaperu.edu.pe

Luis Raymundo
Instituto Peruano de Investigacion en Ingenieria Avanzada
Lima, Perú
lraymundo@bioingenieriaperu.edu.pe

ABSTRACT
Satellite images provide us with information of vital importance, in order to analyze large tracts of land, so their analysis has a degree of complexity characterized by the weight of the image and its size, analyzing large tracts of land originates analyzing the image in all its extension, there are now intelligent algorithms capable of classifying the images, these can analyze the images causing a decrease in the analysis time and improving the result of the analysis of the images. The vegetal cover in our planet is suffering great changes produced by phenomena caused by man, by effects of deforestation, illegal mining among others, that are originating great changes in the terrestrial cover, the evaluation of these changes can be realized by the analysis of satellite images with which you can classify and then locate the area, for this purpose the chromatic characteristics of the images are analyzed with the help of artificial intelligence techniques. In this work, the chromatic characteristics of an image dataset are analyzed. They correspond areas that belong to vegetal cover and areas that do not correspond to the vegetal cover, with the intention of analyzing if these two classes are linearly separable.

CCS Concepts
Theory of computation~Models of learning

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301550

Keywords
Satellite image, segmentation, color bands, image processing, image, characteristics, reflectance, dataset, chromatic, bands.

1. INTRODUCTION
Currently there are many areas of different parts of our planet, is constantly changing, due to the colonizing expansion of man, to exploit the different resources of the and therefore needs to analyze the damage for which they have to analyze for decision-making, many areas that are far from urban areas, it is difficult to assess for the limited access mechanism to these areas, therefore with the advance of technology and with the help of satellites it is possible to capture images that correspond to large extensions of kilometers, the analysis is performed in order to detect if there is damage to land cover.

The presented methodology corresponds to be able to analyze chromatic characteristics of the images that correspond to images with information of the vegetal cover and images that correspond to areas where they have been affected where there is no presence of vegetal cover, the chromatic characteristics will be analyzed: entropy, contrast, homogeneity and energy.

2. MATERIALS AND METHODS
In the present methodology the following steps will be carried out: first we will work with images of a certain area whose coverage is wide, we work with satellite images captured from space with dimensions of 180 kilometers long by 190 kilometers wide, the spatial resolution corresponds to 30 meters, this means that a pixel in the image corresponds to a square of 30 meters from the earth's surface, for the analysis of the images one works with the combination of bands in false composite color, with this image a small dataset was implemented composed of 500 images corresponding to plant cover areas and 400 images corresponding to areas without plant cover that have been damaged, the images of the dataset have a dimension of 25 x 25 pixels, with these images the images were analyzed together by the chromatic characteristics, to determine which of the characteristics presents

a "linearly separable" behavior, this analysis was performed by means of a graph of its values that correspond to each chromatic characteristic "Linearly separable" behavior, this analysis was performed by means of a graph of its values that correspond to each chromatic characteristic.

Below we present a block diagram with processes performed.

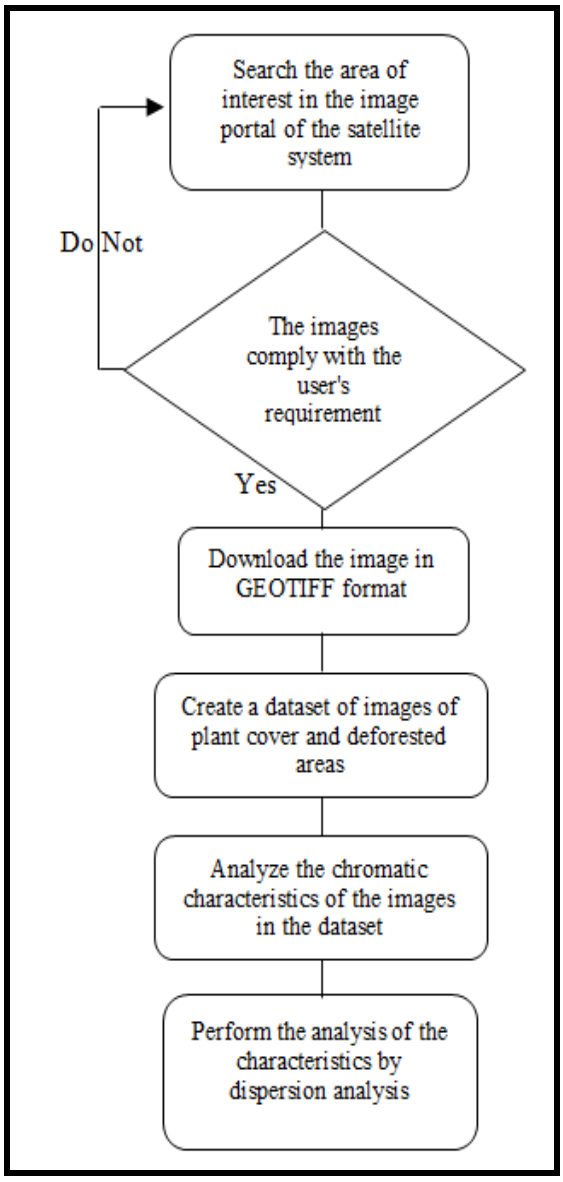


Figure 1.- Description of the presented methodology, presented by means of a flow diagram, indicating each process to be carried out.

We begin with working with the image to be able to evaluate the best combination of bands, in which the vegetal coverage and the non-vegetal coverage can be differentiated, in the following images there are different combinations of bands where the effect of the hand of the man in the zones of forest that corresponds to vegetable cover. We worked with images from the Landsat 8 satellite.

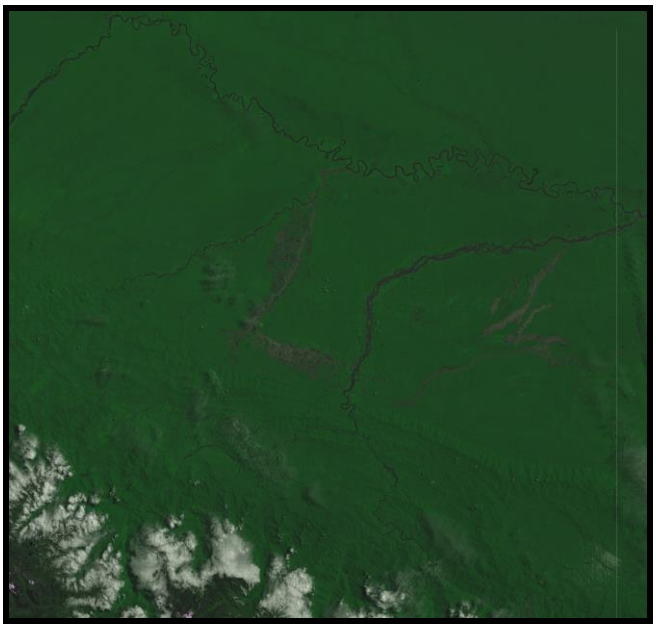


Figure 2. - Image in the combination false color composite.

In the false composite color combination, it can be seen that the areas of green shades correspond to the vegetal cover, the shades of white correspond to the cloud cover and the shades of brown correspond to the deforested areas where it is indicated as coverage non-vegetable.

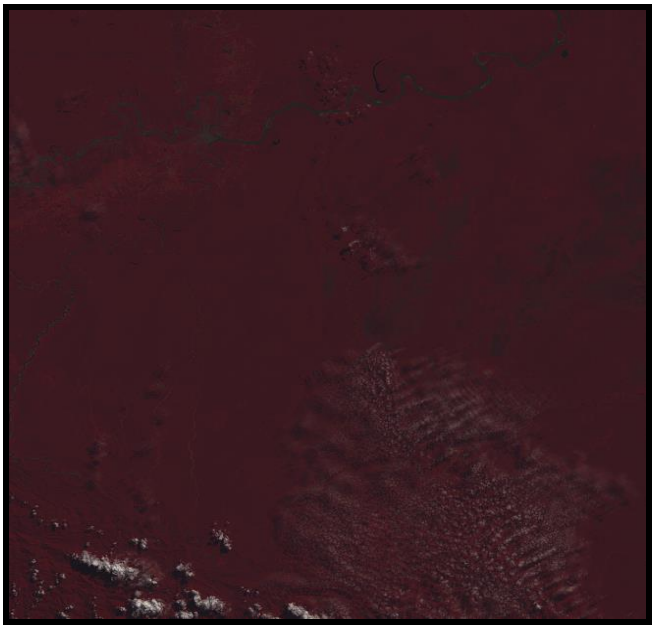


Figure 3. - Image in the real color false combination.

In the real false color combination, it can be seen that the areas of red shades correspond to the vegetal cover, the shades of white correspond to the cloud cover and the black shades correspond to the deforested areas where it is indicated as coverage non-vegetable.

Figure 4. - Image in the real color combination.

In the real color combination, the image is not very clear, it can be seen that the areas of dark tones correspond to the vegetal cover, the tonalities of white correspond to the cloud cover and the yellow tones correspond to the deforested areas where it is indicated as a non-vegetal cover and the bed of the rivers.

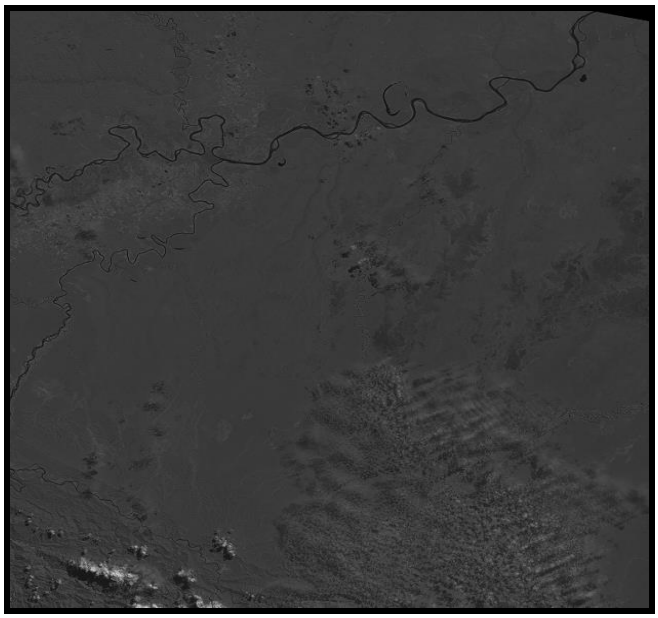


Figure 5. - Panchromatic image

In the panchromatic image, the image is presented in gray scale depending on the resolution of the sensor, it can be seen that the areas of gray tones correspond to the vegetal cover, the tonalities of white correspond to the cloud cover and the tonalities close to the black correspond to deforested areas where it is indicated as non-vegetable cover and river beds.

From these images the one that can best discriminate the areas corresponding to the vegetation cover and the deforested areas that do not correspond to the vegetal cover is the image 2 that corresponds to the combination false color composite.

From the image that corresponds to the composite false color combination, a dataset was implemented with 900 images of which 500 correspond to images of vegetal cover and 400 images that correspond to images that do not correspond to the vegetal cover, these images have a dimension of 25 x 25 pixels, each pixel corresponds to an area of 30 square meters, images of each class are shown in the images presented below.

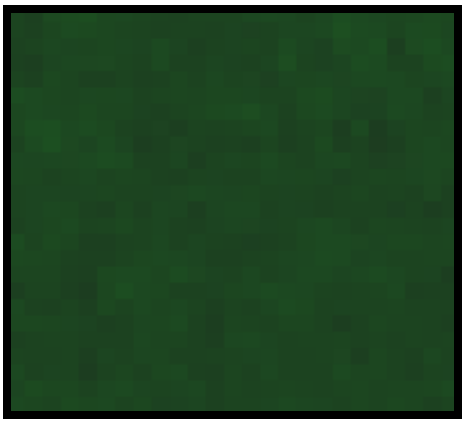


Figure 6. - Image of the dataset that corresponds to the areas of vegetation cover.

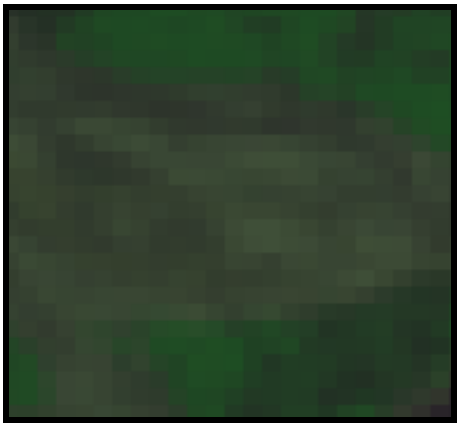


Figure 7. - Image of the dataset that corresponds to deforested areas that do not have plant cover.

After the implementation of the dataset, we proceed to analyze the images by means of chromatic characteristics, which will allow us to obtain a value for each class of the dataset, then we present the characteristics to be evaluated and their respective equation.

CHARACTERISTIC	EQUATION
entropía	$\sum_{i,j}^{N-1} - P_{i,j} \ln(P_{i,j})$
minimal contrast	$\sum_{i,j=0}^{N-1} P_{i,j} \cdot (i-j)^2$
maximum contrast	
minimal homogeneity	$\sum_{i,j=0}^{N-1} i \cdot P_{i,j}$
maximum homogeneity	
minimum energy	$\sum_{i,j}^{N-1} P_{i,j}^{\ 2}$

Figure 8. - Chromatic characteristic with its respective equation.

The texture in an image, according to Emilio Chuvieco, indicates: "That quality refers to the apparent roughness or softness of a region of the image; in short, the spatial contrast between the elements that compose it. The texture of the image comes from the relationship between the size of the objects and the resolution of the sensor. When an object occupies a surface less than 1 mm2 in the image it can not be identified individually, but only through the spatial variability it causes. "(Emilio: 163, 1990).

The entropy, according to Alejandra Pinto, page 21, indicates: "Entropy is used in random measurements (random) of the elements of a matrix, which in this case will be an image. It is the entropy that is responsible for precisely measuring the randomness of the pixels in the co-occurrence matrix. "(Pinto Leal: 21, 2006).

The contrast, according to Alejandra Pinto, indicates: "This concept is totally opposite to the homogeneity where the contrast will have a high value if the high values are concentrated away from the main diagonal and the weight of the probability increases but in quadratic form". (Pinto Leal: 19, 2006).

Homogeneity, according to Alejandra Pinto, indicates: "An image is homogeneous if the values of the main diagonal of co-occurrence matrix are high, this because when observing the homogeneity equation the probability values in the matrix are greater in the diagonal principal and its weight decays exponentially when moving away from the diagonal". (Pinto Leal: 19, 2006).

The energy, according to Alejandra Pinto, indicates: "The values of the co-occurrence matrix are probabilities, so the energy equation maximizes the large values and minimizes the smallest values". (Pinto Leal: 22, 2006).

3. RESULTS

The implementation of the dataset gives us images to be able to evaluate them with the chromatic characteristics, we proceeded to analyze each image of the dataset with each one of the chromatic characteristics by means of its corresponding equation, the value obtained in each one of the images will correspond to the value of the characteristic, this value will be of vital importance to be able to analyze if between the classes of images with vegetal coverage and without vegetal coverage, if there is difference in their values of the evaluated characteristics, it will allow us to use them to construct automatic classifiers for their detection. In the following images, the evaluated characteristics are presented with their respective obtained value.

Vegetable cover

IMAGE		CONTRAST		HOMOGENEITY		ENERGY	
	ENTROPY	minimum	maximum	minimum	maximum	minimum	maximum
	4.0757	2.7723	1.392	0.9325	0.9654	0.7845	0.7889
IMAGE		CONTRAST		HOMOGENEITY		ENERGY	
	ENTROPY	minimum	maximum	minimum	maximum	minimum	maximum
	4.047	2.7692	1.3889	0.9341	0.9669	0.7899	0.7944

Figure 9.- The image presents the evaluated characteristics with their respective value for the images with vegetal cover.

Non-vegetable cover

IMAGE		CONTRAST		HOMOGENEITY		ENERGY	
	ENTROPY	minimum	maximum	minimum	maximum	minimum	maximum
	5.3264	2.9108	0.5278	0.838	0.8975	0.4736	0.5098
IMAGE		CONTRAST		HOMOGENEITY		ENERGY	
	ENTROPY	minimum	maximum	minimum	maximum	minimum	maximum
	5.4023	2.8862	1.4105	0.8756	0.9392	0.5919	0.6437

Figure 10.- The image presents the evaluated characteristics with their respective value for the images without vegetal cover.

4. CONCLUSIONS

The conclusions that are reached when analyzing the images of the dataset with the chromatic characteristics, allow us to determine which of the chromatic characteristics help us to be able to classify areas with vegetal coverage and areas that have been affected by the action of man, these images do not present coverage vegetable, the automatic classifier can be constructed using the characteristics, entropy, homogeneity and energy that were to present a behavior "linearly separable" in view of their values are very distant, this feature greatly helps the classifier to improve its performance.

For the present analysis only the image coming from the combination of color based on "false composite color" was considered. In order to classify we have to work with that combination of bands, in case we want to work with another combination of bands you have to create a new dataset and be able to analyze if the chromatic characteristics present the behavior "linearly separable".

From the analysis of the results and making a visual review of the graphs we can mention in order of priority the chromatic characteristics to work, first the energy characteristic in its minimum variant for presenting a very close values so it becomes uniform its value in the vegetable cover, secondly the characteristic uniformity in any of its variants for presenting constant values in the vegetal cover and thirdly the entropy where its values in the vegetal cover have a slight variation but it differs with the non-vegetal cover.

Then, in order to conclude, the tables are presented with each characteristic analyzed and their respective behavior.

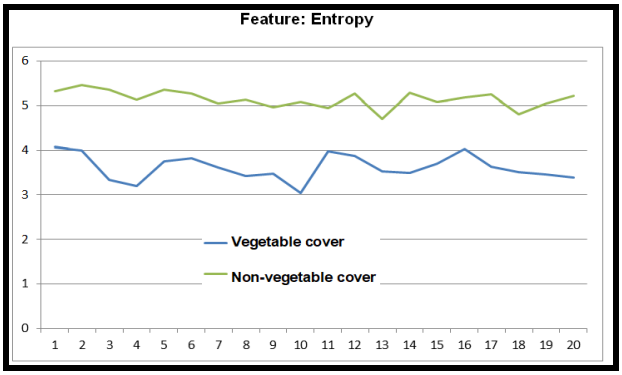


Figure 11.- the entropy characteristic presents a behavior "linearly separable", so it can be used as a relevant feature to classify images with vegetal coverage.

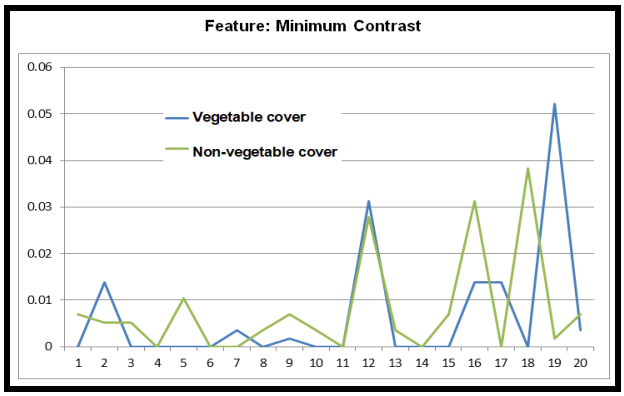


Figure 12.- the characteristic contrast in its minimum variant presents very close values when analyzing the two classes, therefore they do not help to be able to classify them, in this sense it is to be able to perform an automatic classification, this characteristic is discarded.

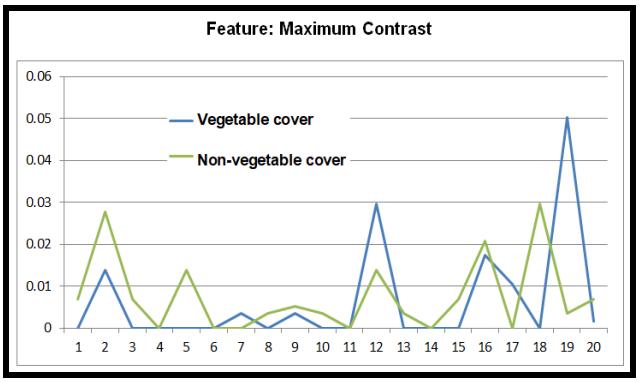


Figure 13.- the characteristic contrast in its maximum variant presents very close values when analyzing the two classes, therefore they do not help to be able to classify them, in this sense it is to be able to perform an automatic classification, this characteristic is discarded.

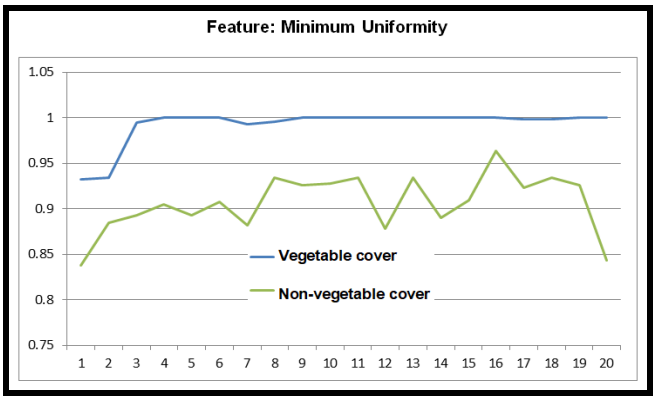


Figure 14.- the characteristic homogeneity in its minimum variant presents distant values therefore it can be used to classify these two classes, in this way this linearity behavior is used to improve the classification.

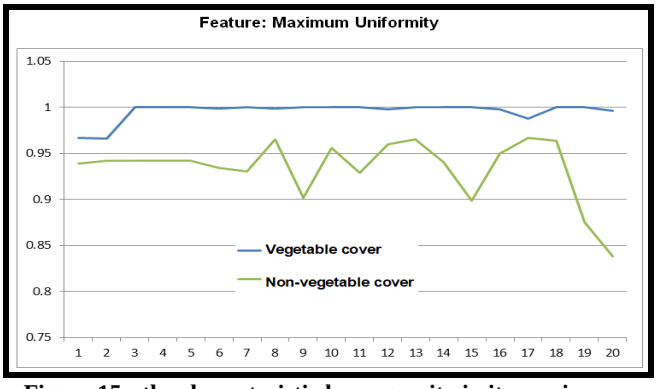


Figure 15.- the characteristic homogeneity in its maximum variant presents distant values therefore it can be used to classify these two classes, in this way this linearity behavior is used to improve the classification.

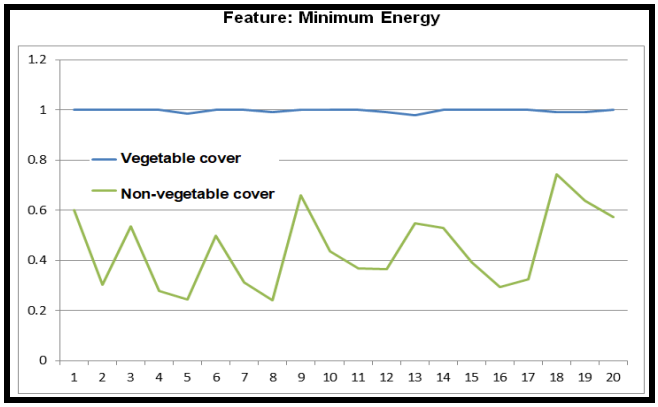


Figure 16.- the energy characteristic in its minimum variant presents distant values therefore it can be used to classify these two classes, in this way this linearity behavior is used to improve the classification.

5. REFERENCES

[1] Lillesand, T., Kiefer, R., Chipman, J., 2004. Remote Sensing and Image Interpretation. fifthed Willey & Sons, New York.

[2] Lunetta, R.S., Lyon, J.G., 2004. Remote Sensing and GIS Accuracy Assessment. CRC press,

[3] Pal, M., Mather, P.M., 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sens. Environ. 86, 554–565.

[4] Russ, J.C., 1999. The Image Processing Handbook. third ed. CRC Press, Boca Raton, FL.

[5] Gonzalez, R. C., & Woods , R. E. (2002). Digital Image Processing, 2nd Edition . Estados Unidos de Amèrica: Prentice Hall.

[6] Emilio, C. (1990). Fundamentos de la Teledetección Espacial. Madrid: Ediciones Rialp S.A. .

[7] Saha, S., Bandyopadhyay, S., 2010. Application of a multiseed-based clustering technique for automatic satellite image segmentation. IEEE Geosci. Remote Sens. Lett. 7, 306–308.

[8] Sridhar, P.N., Surendran, A., Ramana, I.V., 2008. Auto-extraction technique-based digital classification of saltpans and aquaculture plots using satellite data. Int. J. Remote Sens. 29, 313–323.

[9] Pinto Leal, A. C. (2006). Segmentaciòn de imàgenes por textura. Santiago: Universidad de Concepciòn.

[10] Wilkinson, G.G., 2005. Results and implications of a study of fifteen years of satellite image classification experiments. IEEE Trans. Geosci. Remote Sens. 43, 433–440.

[11] Yu, L., Gong, P., 2012. Google Earth as a virtual globe tool for Earth science applications at the global scale: progress and perspectives. Int. J. Remote Sens. 33, 3966–3986.

[12] Romàn Ruiz, T. M. (2013). Clasificación de bosque utilizando imágenes de satélite landsat, con criterio fisiográfico, en la provincia de Maynas, departamento de Loreto Perú. Iquitos: Universidad Nacional de la Amazonia Peruana .

Anchor Graph Hashing with Feature Learning

Weiguang Li
Shenzhen University
Nanshan District, Shenzhen
Guangdong, China
liweiguang2017@email.szu.edu.cn

Xiaogang Peng
Shenzhen University
Nanshan District, Shenzhen
Guangdong, China
pengxg@szu.edu.cn

ABSTRACT
Hash learning methods is a hot topic in the field of image retrieval. Since hash learning methods translate the original datasets into binary codes, they can greatly improve the efficiency of image retrieval. However, the traditional methods cannot well control the information loss. They are not optimal solutions because they learn binary codes in two steps, including low-dimensional feature learning and rotation, so that information loss is not minimal. To address this problem, we integrate dimensionality reduction of datasets and rotation operation into a unified framework. Experimental results prove the promising performance of our method. In experiment section, we select two large-scale datasets and compare with six hash learning methods, the results show our propose method outperforms the other methods.

CCS Concepts
• Information systems→Information retrieval→Speciallized in formation retrieval→Multimedia and multimodal retrieval→I mage search.

Keywords
Image retrieval; hash learning; feature selection.

1. INTRODUCTION
Recently image retrieval on large-scale datasets attracts many attentions. The goal of hash learning is to represent the data samples as a series of fixed-length binary codes (usually using 0/1 or -1/+1 for each of these bits) so that similar samples have similar binary codes, which is an effective way to improve retrieval efficiency and reduce storage space.

In the beginning, *Locality Sensitive Hashing* (LSH) [1][2] is applied widely. It divides the original dataset into multiple subcollections by using the random projections to map datasets, which result the data in each sub-collection is adjacent and the number of elements in the sub-collection is small. However, it requires long binary code for good performance [3][4] [5].

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301535

The local structure of *Spectral Hashing*（SH）[6] is very advantageous in binary representation. Its encoding process of image feature vectors can be regarded as a graph segmentation problem. Besides, the problem can be translated to be a dimensionality reduction problem of Laplace's feature graph by executing spectral analysis on high-dimensional datasets and relaxing constrains, which can obtain hash codes of image data. But SH still has its drawback that the similarity matrix is time-consuming to obtain. And the information loss between real-values of original datasets and binary codes can't be reduced [7].

In order to reduce the information loss, *Iterative Quantization* (ITQ) [8] is proposed. Firstly, it performs PCA dimensionality reduction on the original datasets and then maps the data points to the vertices of a binary hypercube. Secondly, it obtains the rotation matrix iteratively in the purpose of minimizing the corresponding quantization loss. After that ITQ obtain high-quality binary codes for the datasets, which decrease information loss between binary codes and low-dimensional features. However, ITQ has a shortcoming that it focuses on the global information of the datasets so that it cannot preserve the manifold structure of the dataset.

Anchor Graph Hashing (AGH) [4] designed an anchor map for manifold preservation. It uses the neighbor graph between the data clustering center and each data sample point to replaces the original adjacency matrix which greatly reduces the time complexity. In addition, the storage space spent on constructing the map can be decreased because the number of anchor points is much lower than the number of training samples. Although AGH is applied widely, it is still not ideal for controlling the information loss between low-dimensional real values and binary codes.

Large Graph Hashing with Spectral Rotation (LGHSR) [9]is inspired by SH and ITQ. By introducing spectral rotation, the performance is improved consistently. However, it divides the dimensionality reduction and rotation into two steps.

To solve the problem of LGHSR, we combine dimensionality reduction, low-dimensional feature learning and orthogonal rotation in a unified framework, which means optimize the projection matrix and rotation matrix simultaneously and lower the information loss. In the same time, inspired by AGH, the datasets adopt anchor graph for manifold structure preservation. Furthermore, the learned projection matrix is directly used to project the test sample, which will reduce the testing time of the test sample and improve the efficiency of the operation.

All in all, we propose this hash learning method since the other manifold-based hash learning methods mentioned above do not address the problems about information loss and operational efficiency. The contributions of this paper are as follows:

1) Integrating the dimensionality reduction, orthogonal rotation operation and projection learning in one framework that

solve the issue about efficiency caused by dividing data compression and rotation into two steps, such as PCA-ITQ and LGHSR.

2) Using the projection matrix directly to project test sample, we further reduce the time required to generate the binary codes of the test dataset. Information loss can be effectively decreased by our model and good performance can be seen.

We will introduce the construction of the anchor map and the PCA-ITQ algorithm in the second section, explain the construction of our method in the third section, conduct experiments and analyze the results in the fourth section and draw conclusions in the fifth section.

2. RELATED WORKS

In this section, we will introduce the notations and definitions of our model. Then we will explain the construction of the anchor graph and review the hash learning method ITQ.

2.1 Notations and Definitions

We introduce X to represent training data. $X=[x_1,x_2,...,x_i,\ldots,x_n]$, Each of these columns is a d-dimensional feature vector, which constitutes a data matrix $X\in R^{d\times n}$. Then obtain binary codes $B=[b_1,b_2,\ldots b_i,...,b_n]$ through hash learning, where $b^i\in R^l$, l indicates the number of bits. The binary codes learning process can be written as:

$$b_i=sign(Fx_i), \qquad (1)$$

where $F\in R^{l\times d}$ is a encoding matrix， $sign(\square)$ is a binary function， and $sign(Fx_i)$=1 if $Fx_i\geq 0$ and -1 otherwise, which convert real data to binary data.

2.2 Anchor Graph Construction

Manifold learning is an effective way to improve the performance of unsupervised method [10]. However, it takes a lot of computation time to construct the Laplacian graph by using traditional methods. Therefore, we use the anchor to solve the problem of constructing Laplacian graphs of large-scale datasets. The method used in AGH is used here. AGH uses the clustering method to construct the original data set into m anchor points $[u_1,u_2,\ldots u_i,...,u_m]\in R^d$. The elements of the anchor map Z are defined as follows:

$$Z_{ij}=\begin{cases}\dfrac{\exp(-\left\|x_i-u_j\right\|^2/\vartheta)}{\sum_{j'\in\{i\}}\exp(-\left\|x_i-u_{j'}\right\|^2/\vartheta)}, & \forall j\in\{i\}.\\ 0, & otherwise.\end{cases} \qquad (2)$$

where ϑ is a parameter defined in AGH， $\{i\}$ represents the index set of the k nearest neighbor anchor of x_i .

2.3 Iterative Quantization

ITQ obtains binary codes by mapping the original dataset to the hyperplane, which improves the efficiency of image retrieval and reduces the storage space compared to the method of directly comparing image pixels. ITQ minimizes quantization error by iterative rotation matrix operation so that to obtain optimal binary codes. The formula Q for calculating the quantization loss is as follows:

$$Q(B,P)=\left\|B-PV\right\|_F^2 \qquad (3)$$

where $P\in R^{l\times l}$ is a orthogonal rotation matrix， $V\in R^{l\times n}$ indicates datasets after dimensional reduction by PCA. This paper is also inspired by ITQ's formula about quantization errors.

3. JOINTLY LEARNING FRAMEWORK FOR ANCHOR GRAPH HASHING WITH FEATURE LEARNING

3.1 Motivation

Inspired by the AGH anchor map and the ITQ iterative rotation operation, we first cluster the data to construct an anchor map then perform dimensionality reduction and rotation operation synchronously. Finally, the projection matrix will be obtained and be applied directly to the test datasets to cut down the operating time.

3.2 Anchor Graph Hashing with Feature Learning

In order to decrease the information loss between the real value and the binary codes, the proposed model has data compression, rotation matrix, and mapping projection matrix in the iterative algorithm. Our model is designed as follows:

$$\begin{aligned}\min_{B,P,U,V,Q}&\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|b_i-Pv_j\right\|^2 Z_{ji}+\alpha\left\|X-UV\right\|^2+\beta\sum_{i=1}^{m}\left\|b_i-PQ^T\tilde{x}_i\right\|^2\\ &\text{s.t. } b_i\in\{-1,1\}^l,\ P^TP=I,\ U^TU=I\end{aligned} \qquad (4)$$

where b_i is the i-th column data of B , v_j is the i-th column data of V , $\tilde{x}_i$ is the i-th column data of X , $X\in R^{d\times m}$ represents the dataset that forms m anchor points after clustering. $P\in R^{l\times l}$ 、 $Q\in R^{d\times l}$ are projection matrices, $U\in R^{d\times l}$, α 、 β are balance parameters. The whole formula (4) is to minimize the quantization loss during the iterative process. $\left\|b_i-Pv_j\right\|^2$ is used for calculating the minimum quantization error after the rotation operation. $\alpha\left\|X-UV\right\|^2$ aim to perform dimensionality of datasets. $\beta\sum_{i=1}^{m}\left\|b_i-PQ^T\tilde{x}_i\right\|^2$ is used to map the anchor points so that the projection matrix can be used to reduce the dimensionality of testing samples.

3.3 Optimization

We iteratively optimize B,P,U,V,Q to optimize equation (4). The steps to optimize our model are as follows:

Step 1：to calculate the optimal P , given B,U,V,Q ,which are constants. Equation 4 is equivalent to the following:

$$\begin{aligned}&\min_P tr\sum_{i=1}^{m}\sum_{j=1}^{n}(b_i^TZ_{ji}b_i-2b_i^TZ_{ji}Pv_j+v_j^TP^TZ_{ji}Pv_j)\\ &+\beta\sum_{i=1}^{m}tr(b_i^Tb_i-2b_i^TPQ^T\tilde{x}_i+\tilde{x}_i^TQP^TPQ^T\tilde{x}_i) \qquad (5)\\ &\text{s.t. } P^TP=I\end{aligned}$$

Only solve for P , we can deduce the following formula.

$$\max_P tr(PVZB^T+\beta PQ^T\tilde{X}B^T) \qquad (6)$$

Problem (6) has been studied on [7] [11], so we get $(VZB^T+\beta Q^T\tilde{X}B^T)=\tilde{U}\tilde{D}\tilde{V}^T$, so the optimal solution for (6) is：

$$P=\tilde{V}\tilde{U}^T \qquad (7)$$

Step 2：to calculate the optimal U ,given $B,\ P,V,Q$, which are constants. Equation 4 is equivalent to the following:

$$\min_{U} \|X - UV\|^2 \quad \text{s.t. } U^T U = I \qquad (8)$$

Only solve for U , We have $(VX^T) = \hat{U}\hat{D}\hat{V}^T$, so the optimal solution for (8) is：

$$U = \hat{V}\hat{U}^T \qquad (9)$$

Step 3：to calculate the optimal V ,given $B,\ P,U,Q$, which are constants. Equation 4 is equivalent to the following:

$$\min_{V} tr(-2PVZ\tilde{B}^T + VGV^T - 2\alpha X^T UV + \alpha V^T V) \qquad (10)$$

Taking the partial deviation with respect to V to be zero, we can derive the following formula:

$$V = (P^T \tilde{B} Z^T + \alpha U^T X)(G + \alpha I)^{-1}, \quad G_{jj} = \sum_j Z_{ji} \qquad (11)$$

Step 4：to calculate the optimal Q ,given $B,\ P,U,V$. Equation 4 is equivalent to the following:

$$\min_{Q} tr(-2Q^T \tilde{X} B^T P + Q^T \tilde{X} \tilde{X}^T Q) \qquad (12)$$

Taking the partial deviation with respect to Q to be zero, we can d erive the following formula:

$$Q = (\tilde{X}\tilde{X}^T)^{-1} \tilde{X} B^T P \qquad (13)$$

Step 5：to calculate the optimal B , given $Q,\ P,U,V$, Equation 4 is equivalent to the following:

$$\min_{B} tr \sum_{i=1}^{m} \sum_{j=1}^{n} (b_i^T Z_{ji} b_i - 2 b_i^T Z_{ji} P v_j + v_j^T P^T Z_{ji} P v_j) + \beta \sum_{i=1}^{m} tr(b_i^T b_i - 2 b_i^T P Q^T \tilde{x}_i + \tilde{x}_i^T Q P^T P Q^T \tilde{x}_i) \qquad (14)$$

Only solve for B , we can deduce the following formula:

$$\max_{B} tr(PVZB^T + \beta PQ^T \tilde{X} B^T) \qquad (15)$$

We can obtain optimal B by solving the following equation:

$$B = sign(PVZ + \beta PQ^T \tilde{X}) \quad \text{s.t. } b_i \in \{-1, 1\}^l \qquad (16)$$

We summarize the steps of the algorithm in Algorithm 1.

Algorithm 1. Anchor Graph Hashing with Feature Learning

Input: training dataset, testing dataset, the number of anchor point, the number of k nearest neighbors, the maximum number of iterations T, balance parameters α、β , the number of bits.

1: Generate m anchor points by clustering method and generate Z maps using (2)

2: Initialization, B is a random binary matrix, P is an orthogonal m atrix, and Q is an identity matrix.

3：Start iteration.

For $i = 1: T$ do

Step 1: $P = \tilde{V}\tilde{U}^T$

Step 2: $U = \hat{V}\hat{U}^T$.

Step 3: $V = (P^T \tilde{B} Z^T + \alpha U^T X)(G + \alpha I)^{-1}$.

Step 4: $Q = (\tilde{X}\tilde{X}^T)^{-1} \tilde{X} B^T P$.

Step 5: $B = sign(PVZ + \beta PQ^T \tilde{X})$.

4: $F = P \times Q^T$.

Output: Coding matrix F.

4. EXPERIMENTS

In this experiment we selected two datasets CIFAR-10 [12] and MNIST [13] , and then compared our method with the other six hash learning methods, including LSH, SH, AGH, PCA-ITQ, SP, LGHSR. The detailed description of our experiments and analysis of the results as follows.

4.1 Details of Datasets

We chose two large-scale datasets. CIFAR-10 is a dataset with 60,000 images, including 10 tag classes, each image is converted to a GIST feature vector [12] . In the experiment, 59,000 pictures were randomly selected as the training dataset, and the remaining 1000 pictures were used as test samples. Besides, MNIST is a dataset that has 70,000 handwritten digital pictures from numbers 0 to 9 and each picture is translated into a vector with 784 dimensions. 69,000 images were randomly selected as the training set, and the remaining 1000 images were regarded as test sets.

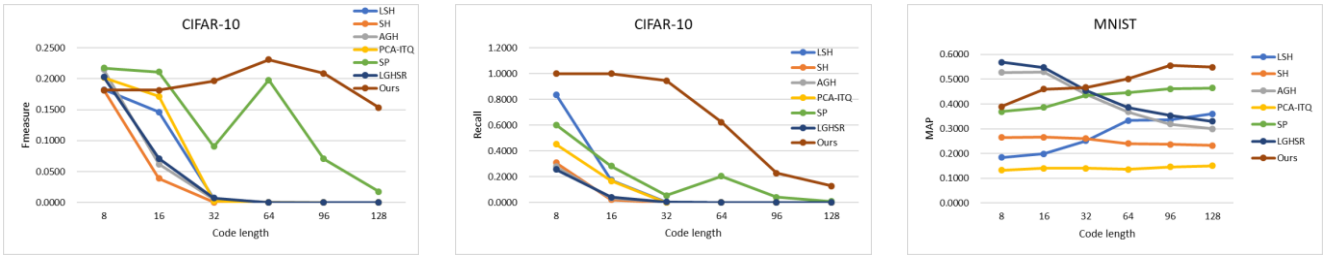
4.2 Experiment Settings

We uniformly set the anchor number AnchorNum=50, the number of k nearest neighbors Knum=5, the maximum number of iterations maxItr=35. Moreover, the parameters of this method α=10^1 , β=10^{-3} , and the number of bits is 8, 16, 32, 64, 96, 128 respectively.

To evaluate the experimental results, we analyze the recall, F-measure, and mean average precision (MAP) with a Hamming radius of 2.

4.3 Analysis

As can be seen from Figure 1, the proposed method has very good results on F-measure, Recall, and MAP. Recall is the highest whatever the length of the binary codes is on the CIFAR-10 dataset. In Figure 1 (a) and 1(c), we can see that when the length of binary codes is not lower than 32, the performance of our method is the best. Because our propose method combines dimensionality reduction and rotation operations, the longer the length of binary codes, the more efficient the iterative operation is.



(a) **(b)** **(c)**

Figure 1. Result of (a)F-measure,(b)Recall of different algorithms on CIFAR-10 dataset and （c）MAP on MNIST dataset.

Table 1. Best Performances of (A)MAP(BITS) and (B)F-MEASURE(BITS) of Different Algorithms on Two Datasets (THE TOP2 ARE EMPHASIZED)

	MNIST		CIFAR-10	
	MAP	F-measure	MAP	F-measure
LSH	0.3599(128)	0.2324(8)	0.1389(96)	0.1825(8)
SH	0.2661(16)	0.3097(8)	0.1266(16)	0.1815(8)
PCA-ITQ	0.4016(128)	0.3276(8)	0.1506(128)	0.2020(8)
AGH	0.5299(16)	**0.5210(8)**	0.1630(16)	0.2145(8)
SP	0.4641(128)	0.3942(8)	0.1577(128)	**0.2174(8)**
LGHSR	**0.5689(8)**	**0.5918(8)**	**0.1668(16)**	0.2030(8)
ours	**0.5553(96)**	0.4121(8)	**0.1727(96)**	**0.2309(64)**

From Table 1, we can see that the MAP and F-measure of our proposed method performs better than other methods on CIFAR-10 dataset. Our method is the top two of the MAP and not much different from the top one on the MNIST dataset. Furthermore, our method has at least 2% improvement compared with the rest of the other methods because the anchor graph of our model can effectively preserve the manifold structure in the high-dimensional dataset.

Although the F-measure is not the top two on the MNIST dataset, but our method is the third one and the difference with the remaining methods is obvious, which can also explain the effect of our method.

5. CONCLUSION

In this paper, we propose a new hash learning model. By experimenting on two datasets and analyzing the experimental results, we conclude that our model is better than the six different hash learning methods listed in the paper. We construct the anchor map, integrate the data compression, rotation operations and mapping matrix learning in an iterative process, which minimize the information loss between datasets and binary codes.

6. REFERENCES

[1] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, “Locality-sensitive hashing scheme based on p-stable distributions,” in *Twentieth Symposium on Computational Geometry*, 2004, pp. 253–262.

[2] A. Andoni and P. Indyk, “Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions,” *Commun. ACM*, vol. 51, no. 1, pp. 117–122, 2008.

[3] W. Liu, C. Mu, and S. Kumar, “Discrete Graph Hashing,” *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2014.

[4] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, “Hashing with graphs,” *Int. Conf. Mach. Learn.*, pp. 1–8, 2011.

[5] F. Shen, C. Shen, Q. Shi, A. van den Hengel, and Z. Tang, “Inductive Hashing on Manifolds,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1562–1569.

[6] Y. Weiss, A. Torralba, and R. Fergus, “Spectral Hashing,” *Nips ’08*, no. 1, pp. 1–8, 2008.

[7] Z. Lai, Y. Chen, J. Wu, W. K. Wong, and F. Shen, “Jointly Sparse Hashing for Image Retrieval,” *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 6147–6158, 2018.

[8] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, “Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, 2013.

[9] X. Li, D. Hu, and F. Nie, “Large Graph Hashing with Spectral Rotation,” *Proc. 31th Conf. Artif. Intell. (AAAI 2017)*, no. 3102015, pp. 2203–2209, 2017.

[10] Y. Chen, Z. Lai, W. K. Wong, L. Shen, and Q. Hu, “Low-Rank Linear Embedding for Image Recognition,” *IEEE Trans. Multimed.*, vol. 20, no. 12, pp. 3212–3222, 2018.

[11] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

[12] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” 2009.

[13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

Robust Weighted Keypoint Matching Algorithm for Image Retrieval

Da-Mi Jeong, Ji-Hae Kim
Sookmyung Women's University
Yongsan-gu, Seoul, Rep. of Korea
+82-2-2077-7293
{dm.jeong,jh.kim}
@ivpl.sookmyung.ac.kr

Young-Woon Lee
Sunmoon University
Asan-si, Rep. of Korea
+82-2-2077-7293
yw.lee@ivpl.sookmyung.ac.kr

Byung-Gyu Kim*
Sookmyung Women's University
Yongsan-gu, Seoul, Rep. of Korea
+82-2-2077-7293
bg.kim@sookmyung.ac.kr*

ABSTRACT
The important factor which controls the accuracy of image matching process is a keypoint extraction. This paper describes a new weighted keypoint matching algorithm for improving the performance of image retrieval. The basic concept is to use the edge information. First, we compute the Scale Invariant Feature Transform (SIFT) features as keypoints from an input image. Also, the edge image is obtained for updating the ranking of the extracted keypoints. With the obtained edge map, we assign weights to keypoints where they correspond on the edge map. Finally the ranking of keypoints is re-ordered and Top-32 keypoints are selected based on the RootSIFT to matching process. Through the experiments, we verify that the proposed algorithm achieves 2.4% of more accuracy than the original SIFT detector when various distortions exist.

CCS Concepts
• **Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems → Interest point and salient region detections**

Keywords
Robust keypoint, Keypoint extraction, Keypoint matching, Image retrieval, Edge information

1. INTRODUCTION
While digital e-books for education are the latest trend, most of educational textbooks are paper based offline publications. These offline textbooks use many works which some people have copyright, such as pictures and photos for educational purpose. However, it is difficult to compensate some fee to the original copyright holders or authors since their works are usually obtained in offline. For a proper compensation, a technique is required to extract the copyrighted works from a publication and detect the copyright holders for the corresponding copyrighted work [1].

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301513

When copyrighted works are used in traditional publication, their properties such as brightness, scale, rotation, and so on are often degraded. Many matching algorithms deal with these transformations, but they usually do not work well with various change or distortion. Among many reported algorithms, Scale Invariant Feature Transform (SIFT) [2] is known as the best one for dealing with rotational translation. The main goal of this paper is to develop more robust keypoint matching algorithm by utilizing the strength of SIFT.

In section 2, we introduce several matching algorithms and compare advantage and disadvantage. The proposed method is explained in section 3. We will describe some experimental results by using our publication dataset in section 4. Finally, the conclusion will be given in section 5.

2. RELATED WORK
The SIFT algorithm is well-known as one of the best keypoint detection algorithm. Although Lowe used Difference of Gaussians (DoG) as an approximation to Laplacian of Gaussians (LoG) so that reduce calculation complexity, the SIFT is still known as very slow since it searches every possible conditions to find keypoints. It uses a corner response function which came from the Harris corner detector [3] to pick out actual keypoints. The Harris corner detector finds a corner if a location has intensity changes through all sides using *E* function of intensity change:

$$E(u,v) = \sum_{x,y} w(x,y)\,[I(x+u, y+v) - I(x,y)]^2, \quad (1)$$

where $w(x, y)$ is window function, $I(x+u, y+v)$ is shifted intensity and $I(x, y)$ is intensity.

This equation is for getting the intensity differences in the places (x, y) moved by (u, v) and (x, y). This equation can simplify like (2) by Taylor expansion. E(u, v) is sum of intensity for movements of (u, v) in x, y, x-y and y-x directions.

$$E(u,v) \approx \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix}, \quad (2)$$

$$M = \sum \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \quad (3)$$

where I_x and I_y are the difference in horizontal and vertical directions in the image respectively and M is a second moment matrix computed from image derivatives.

Than we can define its 'cornerness' by using eigenvalues of matrix M. When both eigenvalues are small, it is flat area. Otherwise, if both are large, the function peaks sharply and it can be the corner. All these equations can simplify one equation.

Equation (4) is the actual corner response function (CRF) as the following:

$$R = \det(M) - K \cdot Trace(M)^2 = \left(I_x^2 \cdot I_y^2 - I_x I_y \cdot I_x I_y\right) - K \cdot (I_x^2 + I_y^2)^2, \quad (4)$$

where K is a pre-defined parameter. We set as K=0.04. If R value is large, the area is most likely to be the corner.

Many similar algorithms have been proposed in order to improve the speed of the SIFT [2]. For instance, Speed-Up Robust Feature (SURF) [4] is well-known algorithm which is faster than the SIFT. Scale Invariant Corner Keypoints (SICK) algorithm [5] has also been reported as a fast keypoint detection scheme. Yang et al. proposed Fast SIFT algorithm [6] which generates an edge group scale space by Sobel edge detector and extracts less keypoints from the edge group space to speed up. Similar to this method, there is also an algorithm to group keypoints on the same outline using the edge information and then to extract affine invariant features [7].

The RootSIFT is another example of the performance improvement while keeping strength of the SIFT [8]. It applies L1 normalization and root operation to descriptors extracted from the SIFT. By using square root (Hellinger) kernel instead of Euclidean distance to define similarity between descriptors, it has higher discrimination so that matching performance can be boosted.

For scale invariance, the SIFT detector uses 4 different scales for one image and each scale composes a pyramid called 'octave' which is applied different Gaussian scale factor. After getting DoG in an octave, a point in middle image is determined the extremum when that point is minimum or maximum among 26 pixels around itself of three neighboring images included itself. The selected extrema may be the actual keypoints, but the extrema obtained by above method may be inaccurate because it is only obtained on the integer domain. Thus, they have to be placed exactly in a continuous space by using Taylor Series. An approximated image has different extremum then binary image. So, if this difference exceeds 50% in either way, this extremum can't be a keypoint because it is closer another location than current location. Furthermore, if the contrast is lower than 0.03, it will not be a keypoint also.

Lastly, only an extremum with enough change of orientation in both vertically and horizontally can be a keypoint. To detect this change, SIFT uses the same idea as the Harris corner detection [3] which is the oldest and well-known interest point detection algorithm, on the other hand, there was also an algorithm that suggested a new measure to extract strong keypoint and 3D surface matching [9].

3. THE PROPOSED METHOD

We are able to suppose there are many robust keypoints around the outline (edge) of an object in the image. The extracted keypoints by the SIFT detector lies on the corner area in many cases. Because keypoints are the extrema which satisfies several constraints of the SIFT detector considering a corner response. If people choose finite number of keypoints in an image, which areas are chosen? Most people will draw keypoints along the outline (edge) in an image.

Hence, the proposed method considers a weight factor to the value of each corner response function of the point corresponds of detected outline (edge). Therefore, the suggested method allows for strong preservation of visual features. We employ the Canny edge detector [10] which becomes known to have better performance with low error ratio while keeping good edge.

The overall procedure of our algorithm is illustrated in Figure 1. The procedure consists of five consecutive parts as the followings:

1. Keypoint extraction by the SIFT detector
2. Canny edge detection
3. Define an edge map by the connected component analysis
4. Weight assignment based on the edge map and update the ranking
5. Computation the RootSIFT by selecting Top-32 keypoints

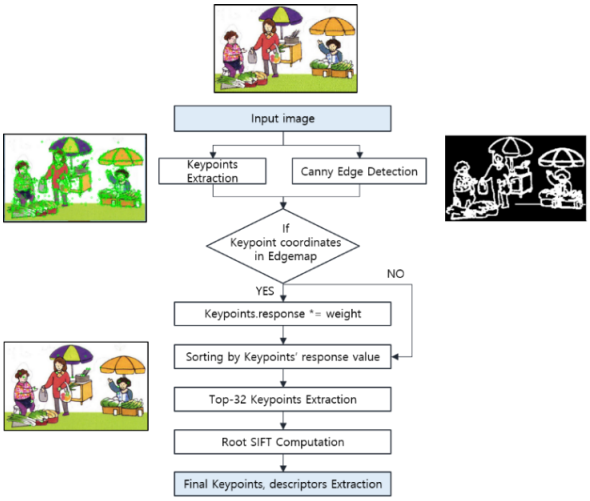


Figure 1. The overall procedure of the proposed method.

3.1 Keypoint Extraction by Using the SIFT Detector

Using the SIFT detector, every keypoints and 128-dimensional descriptors of a keypoint are extracted from the image. Keypoint matching method of the SIFT is to compare the target images with descriptors and give the similarity according to similarity measure. In this study, 128-dimensional descriptor is used in image matching.

3.2 Canny Edge Detection

An input image is blurred by bilateral filtering for removing noises before extracting the edge. Since this filter assigns weights by considering distance and color differences, it can remove noise with leaving edges. Then, we extract the Canny edge from the denoised image. To provide a clear edge, a dilation is applied by 5x5 size of kernel. This kernel size is defined to make edge area wider. The larger the kernel size, the thicker the outline is. We experimentally determine 5x5 size of kernel to fit into images of various size.

3.3 Define An Edge Map by the Connected Component Analysis

How to get coordinates of edge area? Accessing to all pixels consumes too much time depending on the image size. An efficient way to obtain coordinates of edge is to use the connected component analysis [11]. The connected component analysis

combines the connected outline of the image into the same label. Each connected outline is labeled uniquely, which distinguishes its relationship. In this paper, 8 directions-connected component analysis is applied on contour extracted image by the Canny edge detector. Finally, every labeled component are considered as the detected edge map.

Figure 2 shows the process from section 3.2 to 3.3. Figure 2-(a) is example input image. After converting it to the gray scale image, Figure 2-(c) shows the blurred output by a bilateral filter. The outline of this image is extracted shown in (d). Figure 2-(e) shows a result after the connected component analysis. Finally, coordinate information of white outline area of (f) is used as the detected edge map by considering all labeled group as one object.

Figure 2. Process for obtaining the edge map components.

3.4 Weight Assignment Using the Edge Map

If any of keypoint coordinate which is extracted in the previous step, a pre-defined weight is assigned to the response value of the corresponding keypoint where is located on an edge map. The weight function is the following equation:

$$res_{edge}(x_{kp}) = res_{SIFT}(x_{kp}) * \alpha\,, \qquad (5)$$

where res_{edge} is weighted response value of the parameter keypoint, res_{SIFT} is original response value α is a weight. In this paper, we used 1.3 and selected this weight empirically.

3.5 Computation Rootsift by Using Top-32 Keypoints

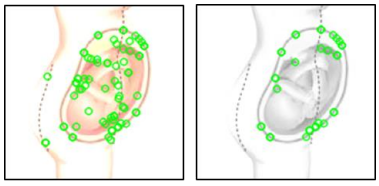


Figure 3. 32 strong SIFT keypoints for matching.

After modifying response values of keypoints based on the edge map, keypoints are sorted according to the rank of these response values. The keypoints which have the higher response values, especially 32 keypoints, are used for re-computing RootSIFT keypoints and descriptor. By using the specific number of strong keypoints, we are able to improve the matching accuracy. If the number of keypoints is less than 32 such as the case of small sized image, some vacant parts are filled with zero values. Figure 3 shows the difference in the number of keypoints. Fig. 3(a) is the keypoints extracted from the original SIFT, and Fig. 3(b) shows the robust keypoints extracted through the edge map computation.

4. EXPERIMENTAL RESULTS

There are several parameters that need to be determined in the proposed method: blurring, dilation kernel size, the number of keypoints to be used for matching, and weight based on the edge map. The limited number of keypoints and kernel size of dilation is determined heuristically. The experiment has been taken into two categories: the edge map weight validation and the retrieval performance for comparison.

For the dataset, the extracted images from scanned files of the offline published books were used. The original images are divided into three groups based on their complexity. Figure 4 shows dataset groups in the experiment. Images of the group A have the size of less 1000px neither a width or a height. The content of image contains very simple so that the background and the foreground can be easily separated. Although the group B has same size to the group A, the group B consists of more complex content and texture. For instance, if an image has several objects or background of complex pattern, it is considered to be a complex image. Lastly, the group C is similar with the group B except that the length of image should be over 1000px, either vertically or horizontally.

Figure 5. Rotated example of the Group A.

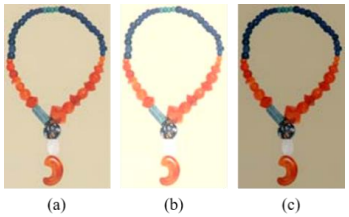


Figure 6. Brightness changes on the original.

Figure 4. Representative images of dataset groups in the experiment.

In order to evaluate the robustness of the proposed algorithm, original images are rotated, brightened and darkened. First, test images were rotated six angels: 15°, -15°, 30°, -30°, 45° and -45° as shown in Figure 5. Also, to give a luminance change, images were made by 10% of brighter or darker. Figure 5(a) shows the original image while Figure 6 (b) and Figure 5 (c) are 10% brighter and darker, respectively.

4.1 Edge Map-based Weight Validation

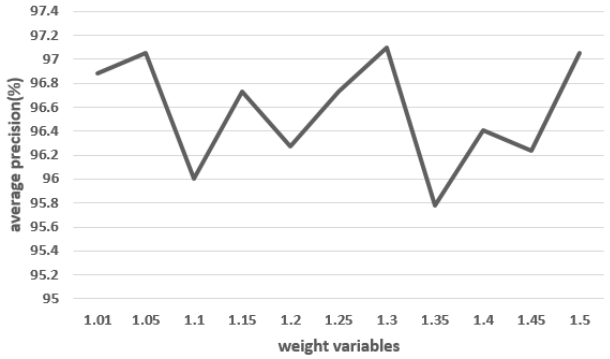


Figure 7. Average precision according to variation of weight α.

The first experiment is to verify which weight can give the best matching result. Figure 7 shows some experiment result about average precisions by weight variables. We took experiments of the transformed dataset with 11 weights between 1.01 and 1.5. From Figure 7, we can observe that the precision with 1.3 as a weight is the highest at 97.09%.

4.2 Performance Analysis

To verify the retrieval performance, we used the precision value as a performance measure in this study. Table 1 shows the performance of different keypoint matching algorithms in terms of the precision. The precision in Table 1 means to average value of the results of all distorted dataset such as rotation and luminance change. The RootSIFT, which is more distinguishable than the SIFT gave higher precision than the original SIFT. However, the result of the group C with complex and larger images had not been much improvement. On the other hand, our method improved the precision by factor of 4.3% and 2.2% in complex image data of the groups B and C, respectively.

Table 1. Precision and the consumed time of the tested algorithms.

Module	Group	Precision (%)	Time (sec)
SIFT [1]	A	97.6	0.51
	B	94.1	0.50
	C	92.5	2.58
	Average	94.7	1.19
RootSIFT [4]	A	97.8	1.06
	B	94.6	1.00
	C	93.0	3.15
	Average	95.1	1.73
Proposed	A	98.3	1.73
	B	98.2	1.70
	C	94.6	4.90
	Average	**97.1**	2.77

5. CONCLUSION

We have proposed a robust image retrieval algorithm based on the weighted keypoint extraction via the SIFT [1]. Based on the outline (edge information) of an object, we assigned a weight value to give the stronger SIFT feature. With the updated response of the SIFT feature, the ranking of the features has been re-ordered. From them, we took Top-32 keypoints for image retrieval. The advantage of the proposed method is to find the stronger keypoints while keeping the strength of the original SIFT algorithm.

We have demonstrated that our algorithm was more accurate in matching than other methods. Through the experiments, the proposed algorithm achieved the more accuracy of the image retrieval by the factor of up to 4.3%.

6. ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A1B04934750) and partially supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2018.

7. REFERENCES

[1] CHOI, Y., J., KIM, J., H., LEE, Y., W., LEE, J.,H., HONG, G., S. AND KIM, B., G. 2017. Efficient Object Classification Scheme for Scanned Educational Book Image. *Journal of Digital Contents Society, 18* (7), 1323-1331

[2] LOWE, D., G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60* (2), 91–110.

[3] HARRIS, C. AND STEPHENS, M. 1988. A combined corner and edge detector. In *Proceedings of the Alvey Vison Conference* (Manchester, UK), Alvey Vision Conference, Organising Committee, 10-5244

[4] BAY, H., ESS, A., TUYTELAARS, T. AND GOOL, V. 2008. Speeded-up robust features (surf). *Computer Vision and Image Understanding 110* (3), 346-359

[5] BO, L., HAIBO, L., ULRIK, S. 2014. Scale-invariant corner keypoints. In *International Conference on Image Processing* (Paris, FRANCE), 5741-5745

[6] YANG, L., LINGSHAN, L., LIANGHAO, W., DONGXIAO, L., MING, Z. 2012. Fast SIFT algorithm based on Sobel edge detector. In *International Conference on Consumer Electronics, Communications and Networks* (Yi Chang, CHINA), 1820-1823

[7] CHENGHO, H., BUI, V., H. 2012. Affine invariant local features based on novel keypoint detection and grouping. In *International Conference on Communications and Electronics* (Hue, VIETNAM), 296-301

[8] ARANDJELOVI, R., ISSERMAN, A. 2012. Three things everyone should know to improve object retrieval. *Conference on Computer Vision and Pattern Recognition* (Providence, USA), Springer-Verlag Berlin Heidelberg, 2911-2918

[9] FENGGUANG, X., XIE, H. 2017. A 3D Surface Matching Method Using Keypoint - Based Covariance Matrix Descriptors. *IEEE Access 5*, 14204-14220

[10] CANNY, J. 1986. A computational approach edge detection. IEEE *Transactions on Pattern Analysis and Machine Intelligence 8* (6), 679-698

[11] HOPCROFT, J. AND TARJAN, R. 1973. Algorithm 447: efficient algorithms for graph manipulation. Communication of the ACM. 16, 6 (Jun. 1973) 372-378. DOI=http://doi.acm.org/ 10.1145/362248.362272

Image Retrieval Method with Fisher GLCM and Graph Model

Honghu Hua
College of Electronic Science, National University of Defense Technology, Changsha, P. R. China
86 13755007744
312841580@qq.com

Jianghua Cheng
College of Electronic Science, National University of Defense Technology, Changsha, P. R. China
86 13874992122
jianghua_cheng@nudt.edu.cn

Tong Liu
College of Electronic Science, National University of Defense Technology, Changsha, P. R. China
86 18570624771
liutong1129@126.com

ABSTRACT
Content-based image retrieval (CBIR) technology is a hot topic for finding the needed images from image big data. A new image retrieval method with fisher GLCM and Graph model is proposed in this paper. The new method designs a new image descriptor with Fisher GLCM features, to extract differentiable features of images; and presents a twice retrieval approach based on Chi-square distance and graph distance by using graph model, to improve the performance of image retrieval. Experiments on ImageCLEF dataset shows that, the new method has highest values of average precision and recall.

CCS Concepts
• **Computing methodologies → Image representations**

Keywords
Image retrieval, Graph model, Fisher, GLCM, CBIR.

1. INTRODUCTION
CBIR technology automatically extract visual features of each image, such as color, texture, shape, then use these features and spatial relationships as indices, to calculate similar distance between query image and target images, finally retrieve relate images from database according to similarity matching [1]. With the development of imaging technology and the popularity of hospital information network, the numbers of medical images for clinical, research and teaching is rapidly expanding. How to easily find the needed image from medical big data has become an important problem currently [2]. For clinical application of medical image retrieval, under undiagnosed condition, diagnosis by means of similar cases from medical big data can bring a lot of help to the doctor, such as improving diagnosis accuracy, reducing troubleshooting time, avoiding the hassle of manually query [3]. Many institutions both at home and abroad are committed to this aspects of research work. Some achievements such as PhotoBook and QBIC system has used in clinical application. It cannot avoid to bring semantic defects in the process of features extraction. Different with normal images, medical images always have high resolution and similar gray distribution. In this situation, it is difficult to use a single global feature to retrieve medical images with same structure or pathology. Local features can deal with medical images better, but have defects such as low efficiency, large difference of different local regions [4,5].

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301521

This paper presents an image retrieval method with fisher GLCM and Graph model. The main contributions include two parts:

(1) This method builds a graph model to improve the performance of image retrieval by twice retrieving according to Chi-square distance and graph distance.

(2) This method presents a new image description method by combining Gray Level Co-occurrence Matrix (GLCM) features, dictionary learning and Fisher encoding, to enhance the robustness to scale and illumination.

2. Graph model
Let $V=\{I_i|i=0,1,\cdots,n\}$ be an image set, where n is the number of images in the set.

Let φ represent an image descriptor for extracting image features, which will be described in the next section.

The feature vector of image I_i can be expressed as

$$v_i=\varphi(I_i) \quad (1)$$

Let D represent a distance measure function used to calculate the distance between two vectors. Commonly used distance measures include Euclidean distance, street distance, Mahalanobis distance, Chi-square distance, and so on. In this paper, we use Chi-square distance as the distance measure. Therefore, the distance between v_i and v_j can be expressed as

$$d_{ij}=\sum\frac{(v_i-v_i)^2}{v_i+v_i} \quad (2)$$

In the process of image query, we calculate the distance between query image and each image in dataset, and then output a query list by the order of distance from smallest to largest, expressed as

$$\tau_q=\left(I_{n_1},I_{n_2},\cdots,I_{n_s}\right) \quad (3)$$

Where, n_s represents the total number of images related to the query image, satisfying the condition that $n_s \ll n$. Therefore, $\tau_q \subset S$.

We use graph model to describe the relationship among images, denoted as

$$G = (V, E) \quad (4)$$

In this model, each image is a vertex. The relationship among images is defined by edge set E.

Let $c(q, n_j)$ represent the relationship measure between image I_q and I_{n_j}. The larger of measure score, the strong relationship of two images.

In this paper, we design the relationship measure based on k-nearest neighbor and distance measure. Details are described as follows.

Let $N_k(q)$ represent an image set containing a given query image I_q and its *k*-nearest neighbors. Let $N_k(q, j)$ represent an image set containing *k*-nearest neighbors of image I_q and I_j. Therefore,

$$N_k(q, j) = N_k(q) \cup N_k(j) \quad (5)$$

The distance between query image I_q and the *i*-th image I_i in $N_k(q, j)$ can be expressed as

$$x_i = \sum \frac{\left(\varphi(I_i) - \varphi(I_q)\right)^2}{\varphi(I_i) + \varphi(I_q)} \quad (6)$$

Similarly, the distance between image I_i and I_j can be expressed as

$$y_i = \sum \frac{\left(\varphi(I_i) - \varphi(I_j)\right)^2}{\varphi(I_i) + \varphi(I_j)} \quad (7)$$

We use cosine measure to design the relationship measure between image I_q and I_j. The score of this relationship measure can be expressed as

$$c(q, j) = \frac{X^{\mathrm{T}} Y}{\sqrt{(X^{\mathrm{T}} X)(Y^{\mathrm{T}} Y)}} \quad (8)$$

Where,

$$X = [x_1, x_2, \cdots, x_{2k}] \quad (9)$$

$$Y = [y_1, y_2, \cdots, y_{2k}] \quad (10)$$

The relationship of two images is actually in proportion to their score of this relationship measure.

When the score exceeds a given threshold t_c, and the test image exists in the query list of query image at the same time, we determine that an edge is exist between image I_q and I_j. Therefore, the edge set can be expressed as

$$E = \left\{ (I_q, I_{n_j}) \middle| \tau_q(n_j) \le n_s, c(q, n_j) \ge t_c \right\} \quad (11)$$

Where, $\tau_q(n_j)$ represent the sort position of image I_{n_j} in the query list τ_q.

3. Image descriptor

In this paper, we extract texture feature of an image, and encode the features by using modified Fisher method to generate image descriptor.

Texture features are suitable features for describing images. *GLCM* is the common texture representation method [6]. It represents the occurrence probability of a pair of pixels that have gray values of *i* and *j* respectively apart *d* distance in θ direction, which can be expressed as $P(i, j|\theta, d)$.

From GLCM we can extract 7 features, including inertia (W_1), energy (W_2), entropy (W_3), uniformity degree (W_4), correlation (W_5), adverse moment (W_6) and variance (W_7), expressed as follows:

$$W_1 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [(i-j)^2 P(i, j | \theta, d)] \quad (12)$$

$$W_2 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [P(i, j | \theta, d)]^2 \quad (13)$$

$$W_3 = -\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [P(i, j | \theta, d) \log P(i, j | \theta, d)] \quad (14)$$

$$W_4 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [P(i, j | \theta, d) / (1 + |i - j|)] \quad (15)$$

$$W_5 = [\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} ijP(i, j | \theta, d) - \mu_x \mu_y] / \sigma_x \sigma_y \quad (16)$$

$$W_6 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [P(i, j | \theta, d) / (1 + (i-j)^2)] \quad (17)$$

$$W_7 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [(i-m)^2 \log P(i, j | \theta, d)] \quad (18)$$

Where,

$$\mu_x = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} iP(i, j | \theta, d) \quad (19)$$

$$\mu_y = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} jP(i, j | \theta, d) \quad (20)$$

$$\sigma_x^2 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i - \mu_x)^2 P(i, j | \theta, d) \quad (21)$$

$$\sigma_y^2 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (j - \mu_y)^2 P(i, j | \theta, d) \quad (22)$$

Where *m* is the average value of $P(i, j|\theta, d)$, *L* is the gray level of the image.

The features are different with different parameters θ and *d*. In this paper, we set these parameters as $\theta = \{0^o, 45^o, 90^o, 135^o\}$ and $d = \{1, 2\}$. In this way, we can extract 56 (=7×4×2) GLCM features of every image blocks.

Standard Bag of Word (BoW) model represents an image by applying Fisher framework in visual word index [7]. Let

$\pi=\{\pi_1, \pi_2,\cdots,\pi_k\}$ represent the polynomial learned from *K* visual words, where π_k can be expressed as a logarithmic form:

$$\pi_k=\frac{e^{-\gamma_k}}{\sum_{i=1}^{K}e^{-\gamma_i}} \quad (23)$$

Where, γ_i represents the weight coefficient of the *i*-th visual words.

In this way, the similarity based on BoW model can be expressed as

$$p(w_{1:N})=\prod_{i=1}^{N}p(w_i=k) \quad (24)$$

Where,

$$p(w_i=k)=\pi_k \quad (25)$$

In order to reduce the complexity, we use logarithmic form to calculate the gradient of log-likelihood ratio to weight coefficient, expressed as

$$\frac{\partial\sum_{i=1}^{N}\ln p(w_i)}{\partial\gamma_k}=n_k-N\pi_k \quad (26)$$

Where, n_k represents the number of times of visual Word *k* in $w_{1:N}$.

For any two different images, there must be a different π_k between two models after modeling based on BoW model. In terms of this thought, we use Dirichlet function to model $\pi=\{\pi_1, \pi_2,\cdots,\pi_k\}$, in order to fully exploit the relationship between visual word index. The modeling formula can be expressed as

$$p(\pi)=D(\pi|\alpha) \quad (27)$$

Where, $D(\cdot)$ represents the Dirichlet function. $\alpha=\{\alpha_i|i=1,2,\cdots,K\}$, where α_i indicating the coefficient of the *i*-th Visual Word index. In this paper, we set *K*=1000.

Based on this model, the edge distribution on $w_{1:N}$ can be expressed as

$$p(w_{1:N})=\int p(\pi)\prod_{i=1}^{N}p(w_i|\pi)d\pi \quad (28)$$

We use Gamma function $\Gamma(\cdot)$ to convert edge distribution to integer form, expressed as

$$p(w_{1:N})=\frac{\Gamma(\hat{\alpha})}{\Gamma(N+\hat{\alpha})}\prod_{k=1}^{K}\frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)} \quad (29)$$

Where,

$$\hat{\alpha}=\sum_{k=1}^{K}\alpha_k \quad (30)$$

According to Fisher's framework, we use the gradient of log-likelihood ratio of visual word to hyper-parameters to represent the image, expressed as

$$\frac{\partial\ln p(w_{1:N})}{\partial\alpha_k}=\psi(n_k+\alpha_k)+\psi(\hat{\alpha})-\psi(N+\hat{\alpha})-\psi(\alpha_k) \quad (31)$$

Where,

$$\psi(x)=\frac{\partial\ln\Gamma(x)}{\partial x} \quad (32)$$

The posterior probability distribution can be expressed as

$$p(w=k|w_{1:N})=\int p(w=k|\pi)\,p(\pi|w_{1:N})d\pi \quad (33)$$

Taking into account the uneven brightness and various texture of medical images, we execute image partitioning to divide the image into several blocks. Firstly, let b1 represent the whole image, used to extract global features. Then, we use quadtree partitioning to get four blocks denoted as b2~b5. And then, we go on this partitioning to get blocks b6~b21 from b2~b5 and b22~b85 from b6~b21. In this way, an image is divided into 85 image blocks. We will extract features from these blocks to describe the global and local properties of the image. Taking into account the texture features of 85 blocks are independent, we are unable to use Fisher kernel to encode GLCM features directly. Therefore, we extend the above BoW model. The basic idea is: using *T* groups of hybrid models to model the visual words in an image. Groups **T** is same with block number, namely *T*=85. We use polynomial θ to encode *T* groups. Every member of each group (that is GLCM features of each block) still use polynomial distribution π of *K* visual words. Thus,

$$p(w_i=k|\theta,\pi)=\sum_{t=1}^{T}p(z_i=t|\theta)\,p(w_i=k|\pi_t) \quad (34)$$

The marginal distribution on $w_{1:N}$ can be expressed as

$$p(w_{1:N})=\iint p(\theta)\,p(\pi)\prod_{i=1}^{N}p(w_i|\theta,\pi)d\theta d\pi \quad (35)$$

By combining with Dirichlet prior, we can get

$$p(\theta)=D(\theta|\pi) \quad (36)$$

By introducing hyper-parameters, the distribution can be expressed as

$$p(\pi_t)=D(\pi_t|\eta_t) \quad (37)$$

Thus, a posteriori probability can be expressed as

$$q(\theta,\pi_{1:T},z_{1:N})=q(\theta)\prod_{i=1}^{T}q(\pi_t)\prod_{i=1}^{N}q(z_i) \quad (38)$$

Where,

$$q(\theta)=D(\theta|\alpha^*) \quad (39)$$

$$q(\pi_t)=D(\pi_t|\eta_t^*) \quad (40)$$

$$\alpha_t^*=\alpha_t+\sum_{i=1}^{N}q(z_i=t) \quad (41)$$

$$\eta_{tk}^{*} = \eta_{tk} + \sum_{i=1, w_i=k}^{N} q(z_i = t) \quad (42)$$

The above parameters can be calculated in the iteration process.

Two gradients of log-likelihood ratio to two hyper-parameters can be expressed as

$$\frac{\partial \ln p(w_{1:N})}{\partial \alpha_k} = \psi(\alpha_k^{*}) + \psi(\hat{\alpha}) - \psi(\hat{\alpha}^{*}) - \psi(\alpha_k) \quad (43)$$

$$\frac{\partial \ln p(w_{1:N})}{\partial \eta_{tk}} = \psi(\eta_{tk}^{*}) + \psi(\hat{\eta}_t) - \psi(\hat{\eta}_t^{*}) - \psi(\eta_{tk}) \quad (44)$$

In the end, we use the gradients of each visual word to generate a Fisher vector as the image descriptor φ.

4. Image retrieval

In this paper, we execute image retrieval based on graph model and Fisher GLCM features. The basic idea is that: using traditional Chi-square distance for first time retrieval, to narrow the scope of image retrieval; then using graph distance for second time retrieval, to reduce false positive phenomena. The training stage and image retrieval stage are two main stages for our image retrieval method.

In training stage, for each image in the dataset, we extract feature vector $v_i = \varphi(I_i)$ according to the image descriptor, and calculate and save the Chi-square distance of two feature vectors corresponding to any two images.

In image retrieval stage, for a query image I_q, we also extract feature vector $v_q = \varphi(I_q)$ according to the image descriptor shown in Section 2, and calculate the Chi-square distance d_{qi} between its feature vector and the one of any image I_i in the dataset.

Then, we get the query list $\tau_q = (I_{n_1}, I_{n_2}, \cdots, I_{n_s})$ by the order of distance from smallest to largest, in this way to complete the first time retrieval.

And then, giving an initial threshold $t_c = t_s$, we calculate the relationship $c(q, j)$ between I_q and any image I_j in τ_q, to build a graph model $G_1 = (V, E_1)$ as described in Section 3. Where, V contains all images in dataset and query image. In this paper, t_s =0.5.

We repeat the above steps by increasing threshold $t_c = t_c + \Delta t$, continue to build graph model $G_i = (V, E_i)$. Where, index i represents the i–th graph model. By repeating this process until $t_c \geq 1$, we can get the final set of graph models, denoted as $G_{all} = \{G_i = (V, E_i) | i = 0,1,\cdots,m\}$. Where, m represents the total number of graph models built above, Δt =0.02.

We calculate the graph distance between two graph models. The more similar of two images, the stronger relationship of them. Therefore, graph distance can be measured by the number of edges at different threshold t_c. Let d'_{qn_j} be the graph distance between query image I_q and any image I_{n_j}. The initial value of d'_{qn_j} is set to m. Traversing each graph model in G_{all}, if there is an edge between query image I_q and any image I_{n_j} in a graph model, $d'_{qn_j} = d'_{qn_j}$ -1. After traversing all graph model, we get the final graph distance between query image I_q and any image I_{n_j}.

Finally, we execute second time retrieval according to list sorted by graph distance from smallest to largest, and get the final query list $\tilde{\tau}_q$. In the end, we output T images in front of $\tilde{\tau}_q$ as the final retrieval results, where T represents the number of images we want return. In this paper, T=30.

5. Experiment and Analysis

We select ImageCLEF dataset [8] as experimental dataset, to test the performance of image retrieval method. This ImageCLEF image dataset contains many X-ray images of patients with different age and gender. The brightness and contrast of images in this dataset vary greatly. In order to ensure that the image number of each type roughly the same, we select 4471 images from this dataset, which contains 1639 query images and 50 categories image dataset for retrieval.

There are two commonly used evaluation metrics for image retrieval: average precision (Pre) average and recall (Rec). The higher value of Pre and Rec, the better performance of image retrieval.

In the rest of this section, we first analysis the influence of some parameters in our methods, then compare the performance of different image retrieval methods.

5.1 Parameters analysis

Parameters n_s and k in our method make great impact on image retrieval. Figure 1 and Figure 2 show the image retrieval results of our method with different values of n_s and k on the ImageCLEF dataset.

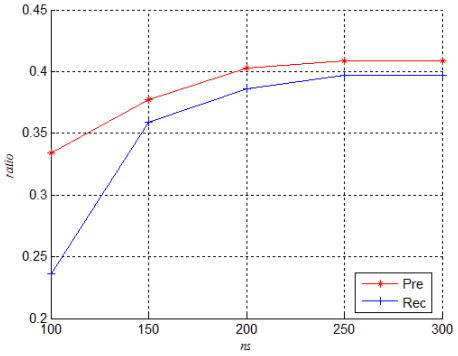


Figure 1. The results with different values of n_s

In Figure 1, we set k=15. From the curves in Figure 2 we can see, the values of Pre and Rec increase along with the value of n_s. While n_s >250, the values of Pre and Rec become stable. The larger of n_s, the lower of image retrieval efficiency. Therefore, we set n_s = 250.

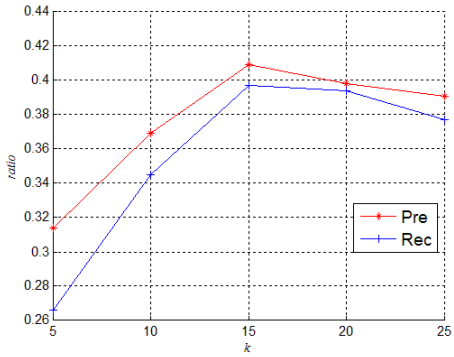


Figure 2. The results with different values of *k*

In Figure 2, parameter k varies but n_s = 250. The values of Pre and Rec reach peak while k=15. Therefore, we set k=15.

5.2 Performance comparison

This section we execute image retrieval experiments on the ImageCLEF dataset by using different image retrieval methods, to compare the proposed method with the current image retrieval methods described in reference [9] and [10]. The curves of average precision and recall with different methods under same conditions of image retrieval are shown in Figure 3.

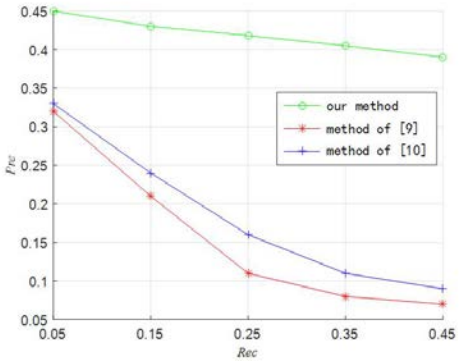


Figure 3. The curves of average precision and recall

It can be seen obviously form Figure 3 that, the values of Pre and Rec of our method are higher than the other two methods at the same situation. More importantly, the Pre value of our method decreases along with Rec value more slowly, while the downward trend of the other two methods is obvious. It means that, the image retrieval performance of our method is more stable. The reason is that, our method reduces false positive phenomena and thereby improves image retrieval performance by means of twice retrieval with feature distance and graph distance.

6. Conclusions

This paper presents an image retrieval method with fisher GLCM and Graph model. In terms of the properties of images, this method designs a new image descriptor combined with GLCM features and Fisher coding, and presents a twice retrieval approach based on Chi-square distance and graph distance, to improve the performance of image retrieval. This method is an effective image retrieval method, which can help for exploitation of image big data.

7. ACKNOWLEDGMENTS

The authors would like to thank all the reviewers for their valuable suggestions and comments. This work was supported by the National Natural Science Foundation of China under Grant number 41301491.

8. REFERENCES

[1] Feng L, Liu S, Xiao Y, et al. A novel CBIR system with *WLLTSA* and ULRGA[J]. Neurocomputing, 2015, 147(1):509-522. DOI= https://doi.org/10.1016/j.neucom.2014.06.027.

[2] Tzanis G. Biological and Medical Big Data Mining[J]. International Journal of Knowledge Discovery in Bioinformatics, 2014, 4(1):42-56. DOI= https://doi.org/10.4018/ijkdb.2014010104.

[3] Gaithayadi M, Bouslimi R, Akaichi J. A New CBIR Approach for the *Annotation* of Images[J]. International Journal of Computer Applications, 2014, 44(73):34-45. DOI= https://doi.org/10.5120/12747-9670.

[4] Tarjoman M, Fatemizadeh E, Badie K. An implementation of a CBIR system based on SVM learning scheme[J]. Journal of Medical Engineering & Technology, 2013, 37(1):43-47. DOI= https://doi.org/10.3109/03091902.2012.742157.

[5] Zhang D, Lu G. Review of shape representation and description techniques[J]. Pattern Recognition, 2004, 37(1):1-19. DOI= https://doi.org/10.1016/j.patcog.2003.07.008.

[6] Arebey M, Hannan M A, Begum R A, et al. CBIR for an *automated* solid waste bin level detection system using GLCM[J]. Lecture Notes in Computer Science, 2011, 7066:280-288. DOI= https://doi.org/10.1007/978-3-642-25191-7_27.

[7] Akaichi J. Bag of words for semantic automatic image annotation[J]. Network Modeling Analysis in Health Informatics and Bioinformatics, 2014, 3(1):1-7. DOI= https://doi.org/10.1007/s13721-014-0061-2.

[8] *ImageCLEF* - The CLEF Cross Language Image Retrieval Track. http://www.imageclef.org/.

[9] Kumar A, Nette F, Klein K, et al. A Visual Analytics Approach Using the Exploration of Multidimensional Feature Spaces for *Content*-Based Medical Image Retrieval.[J]. IEEE Journal of Biomedical & Health Informatics, 2014, 19(5):1734-1746. DOI= https://doi.org/10.1109/JBHI.2014.2361318.

[10] Bugatti P H, Kaster D S, Ponciano-Silva M, et al. PRoSPer: Perceptual similarity queries in medical CBIR systems through user profiles[J]. Computers in Biology & Medicine, 2014, 45(45C):8-19. DOI= https://doi.org/10.1016/j.compbiomed.2013.11.015.

Chapter 5: Video Processing Technology and Method

A Long-Short Term Memory Neural Network Based Rate Control Method for Video Coding

Zheng-Teng Zhang
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
zhang_zhengteng@dahuatech.com

Jucai Lin
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
lin_jucai@dahuatech.com

Ruidong Fang
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
fang_ruidong@dahuatech.com

Juan Lu
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
lu_juan@dahuatech.com

Yao Chen
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
chen_yao@dahuatech.com

ABSTRACT
In industrial and practical application, the robust rate control method is highly required in field of video coding. In this paper, a Long-Short Term Memory (LSTM)-based method is proposed to optimize x264 rate control. The bitrate and distortion information are used to construct the LSTM recurrent neural network to predict a quantization parameter (QP) at inter frame level. Then a QP refinement strategy is utilized to further improve the relationship between the QP and bitrate. Finally, the average bitrate control (ABR) method is optimized to provide more accurate bitrate while encoding. Experimental results show that the proposed method achieves 1.2% BD-rate reduction in HM test sequence, and 1.0% in surveillance video. Meanwhile, the target bit matching rate is up to 98.9% and 99.7%, respectively.

CCS Concepts
• Information systems→ Multimedia streaming • Computing methodologies→ Neural networks. • Computing methodologies→ Image compression.

Keywords
x264; video coding; rate control; Long-Short Term Memory Neural Network (LSTM);

1. INTRODUCTION
Video coding technique play a significant role in high resolution video storage and transmissions, such as UHDTV, VR (visual reality) and surveillance video. The H.264 [1] and HEVC (High Efficiency Video Coding) [2] are the most popular video coding standard focuses on optimizing the bitrate saving for high resolution video. The rate control provides a specific bitrate coding which is closed to the target bitrate in video coding. In order to achieve more stable and accurate bitrate keeping visual quality of video, the better rate control method is highly required in industrial and practical application.

The rate control method in H.264 manages bits based on quadratic R-D model, which is released by JVT-G012 (joint video team) [3]. Firstly, the bit allocation is employed in GOP, frame and basic unit level, respectively. Then the quadratic R-D model is utilized to predict quantization step with the temporal distortion. The quantization parameter (QP) and langrage multiplier λ are eventually decided for encoding and rate distortion optimization (RDO), respectively. The mentioned above method is well solved the "chicken-and-egg" legacy and further extended to HEVC rate control, i.e. JCTVC-H0213 (joint collaborative team on video coding) [4], which use the pixel-wise unified rate quantization (URQ) model to replace the typical quadratic R-D model for rate control [5]. However, the method cannot satisfy the bitrate matching with complexity coding technique in HEVC. Thus the λ-domain rate control method, released by JCTVC-K0103 [6], is proposed to optimize both RDO and rate control [7].

x264 is an open source encoder based on H.264 standard [8]. The rate control method in x264 is different to H.264 reference software JM. There are two different rate control methods at x264 frame level, the average bitrate control (ABR) method and constant rate factor (CRF) method. The CRF method is utilized to provide a constant visual quality for each frame while encoding. It means that the bits consuming is not considered in this method. The ABR method is triggered to provide the actual encoded bitrate which is closed to the target bitrate in any period of encoding time. However, the ABR method does not let the bit consuming match the target bits perfectly in each frame. The bit allocation for each frame is roughly and unstable. In this paper, the Long-Short Term Memory neural network (LSTM) is thereby proposed to improve the bitrate accuracy and coding performance.

The reminder of this paper is organized as follows. The background review is introduced in Section 2, including the

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301524

review of deep learning-based method in video coding and the original ABR method in x264 rate control. In Section 3, the proposed LSTM network is given a statement. Then, Section 4 clarifies the proposed LSTM-based rate control method in x264. The experimental result and details of implementation are introduced in Section 5. Finally, the conclusion and future work is stated in Section 6.

2. BACKGROUND REVIEW

2.1 Video Coding and Deep Learning

Recently, deep learning is widely used in field of video coding. In general, the deep learning approach solves the video coding problem in two categories, visual quality improvement and coding complexity reduction. According to subjective or objective quality improvement, Dai et al. [9] proposed a fully convolutional neural network (CNN) for video post processing. The reconstructed frame can be further improved its visual quality by the pre-train CNN model. For complexity reduction, Xu et al. [10] proposed a deep learning-based fast mode decision algorithm for CTU. The CU partition is predicted by an early-terminated hierarchical CNN (ETH-CNN) in HEVC intra-mode. Meanwhile, in inter-mode, the temporal correlation of the CU partition is solved by ETH-LSTM, which combines the CNN and LSTM structure. Furthermore, the traditional and advanced machine learning technologies are also applied in rate control. A joint support vector machine (SVM) and game theory model is proposed for HEVC rate control at inter frame CTU level [11]. Li et al. [12] proposed a CNN based approach to predict the model parameters in R-λ model for HEVC rate control. However, the existing rate control methods do not apply deep learning model which can present strong temporal correlation, i.e. LSTM.

The LSTM network can well perform the temporal correlation for each frame in video. Donahue et al. [13] proposed long-term recurrent convolutional networks (LRCNs), a class of architectures for visual recognition and description which combines convolutional layers and LSTM layers. The temporal correlation also exists in rate control. The bitrate and distortion information in previous frames are highly related to the quantization parameter (QP) of current frame. Therefore, the LSTM-based rate control method is proposed in this paper.

2.2 The ABR Rate Control Method in x264

The ABR method in x264 depends on distortion and bitrate information to predict QP. The sum of absolute transformed difference (SATD) becomes as distortion information in pre-analysis process. The target bits and the coded bits from previous frame are the bitrate information. Firstly, a parameter called *qscale* is calculated based on SATD. Then the initial qscale (rate control equation qscale, *rceq*) is further modified based on coded bits and target bits, respectively. It exists a *qscale*-QP model that let the final *qscale* map to QP for frame encoding. The details of calculation can be formulated as:

$$rceq[i] = \left(\frac{\sum_{j=0}^{i}(0.5^{i-j}\cdot SATD[j])}{\sum_{j=0}^{i}0.5^{i-j}}\right)^{1-qcompress} \quad (1)$$

$$qscale[i] = \frac{rceq[i]}{B_{wanted}}\cdot\sum_{j=0}^{i-1}\frac{qscale[j]\cdot B_{coded}}{rceq[j]} \quad (2)$$

$$qscale[i] = qscale[i]\cdot overflow \quad (3)$$

where *i* is the number of the encoded frame; *qcompress* is a constant value calculate the *rceq*[i]; The *SATD*[j] means the SATD value at *j* frame, which is extracted from the residual difference between the original frame and predicted frame. The predicted frame is constructed by the original frame and last reference frame during the motion compensation, which is located at pre-analysis processing before the frame encoding. In equation (3), the overflow can be expressed as:

$$overflow = 1+\frac{B_{total}-B_{wanted}}{2t\cdot R_{target}\cdot\max(1,\sqrt{\frac{i}{fps}})} \quad (4)$$

where B_{total}, B_{wanted} and B_{coded} are the total encoded bits, total target bits and encoded bits of previous frame through current encoding frame, respectively. R_{target} is the target bitrate. t is set to 1 in default, as a constant value which means the tolerance for measuring the fluctuation of *qscale*. Finally, the *qscale*-QP model is given as:

$$QP[i] = a + m\cdot lb\left(\frac{qscale[i]}{n}\right) \quad (5)$$

where *a*, *m* and *n* are the model parameters and set to 12, 6 and 0.85, respectively. $lb(\cdot)$ is log based binary.

3. PROPOSED LSTM NETWORK FOR RATE CONTROL

3.1 The Rate Control based LSTM

As aforementioned, the QP generated in x264 depends on the SATD value and coded bits information which are from the current and previous frame. The temporal correlation is proper to generate a LSTM network to predict QP. Therefore, the following parameters in list are become as an input features $x \in R^L$ to construct the proposed network with *L*=6 neurons:

x_1: The SATD per pixel of current frame.
x_2: The mean square error (MSE) of last encoded frame.
x_3: The structure similarity index (SSIM) of last encoded frame.
x_4: The coded bits per pixel of current frame.
x_5: The coded bits per pixel of last encoded frame.
x_6: The encoded QP of last frame.

All of the input features can be extracted during the x264 encoding. The distortion and bitrate information are considered with temporal correlation. Thus the proposed rate control LSTM (RCLSTM) network is presented at Fig. 1. An input feature sequence x_L^T with time step *T* corresponds to a training sample extracted input features from P frames in x264 encoding. In general, the input sequence $\langle x_L^1, x_L^2, \ldots, x_L^T\rangle$ has *T* length time step corresponding to the *T* continuous frames in video.

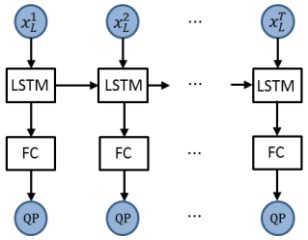


Figure. 1 The proposed RCLSTM network

Table 1. The architecture of RCLSTM network

Layers	Input Features	LSTM+ ReLU	FC+ReLU	FC
Neurons	6	128	128	1
Size	$L\times T$	$128\times T$	$128\times T$	$1\times T$

3.2 Proposed LSTM Architecture

Table 1 shows the details of proposed RCLSTM network. For the LSTM layer, let T as the time step of the input features x_L^T. The forward processing of LSTM are:

$$i^T = \sigma(W_i \cdot [x_L^T,\ h^{T-1}] + b_i) \quad (6)$$
$$f^T = \sigma(W_f \cdot [x_L^T,\ h^{T-1}] + b_f) \quad (7)$$
$$o^T = \sigma(W_o \cdot [x_L^T,\ h^{T-1}] + b_o) \quad (8)$$
$$g^T = \tanh(W_c \cdot [x_L^T,\ h^{T-1}] + b_c) \quad (9)$$
$$c^T = i^T \odot g^T + f^T \odot c^{T-1} \quad (10)$$
$$h^T = o^T \odot \tanh(c^T) \quad (11)$$

where $\sigma(\cdot)$ and $\tanh(\cdot)$ donate the element-wise non-linearity, corresponding to sigmoid and hyperbolic tangent function; and $\odot$ donates element-wise multiplication. i^T, f^T and o^T are the input gate, forget gate, and output gate in LSTM unit, respectively. W_i, W_f, W_o and W_c are trainable weights for LSTM neurons; and b_i, b_f, b_o and b_c are their biases. The output h^T is continuously used in the following fully connected (FC) layer, given as:

$$F_2(h^T) = max(0, W_2 \cdot F_1(h^T) + b_2) \quad (12)$$
$$F_3(h^T) = W_3 \cdot F_2(h^T) + b_3 \quad (13)$$

where W_2 and W_3 are the weight parameters of the FC layer, and b_2 and b_3 are their biases. $F_1(h^T)$ is the output of the LSTM layer with the length of time step T, expressed as:

$$F_1(h^T) = max(0, h^T) \quad (14)$$

where $max(\cdot)$ donates the rectified linear unit (ReLU) function. The final output $F_3(h^T)$ is the predicted QPs which we expected for x264 frame encoding. In order to achieve the accurate QP value, the Euclidean loss function is employed at training stage:

$$loss = \frac{1}{N}\sum_{n=1}^{N} \left\| F_3(h^T)_n - \mathrm{y}^T{}_n \right\|^2, \quad (15)$$

where N is the total number of training samples. $\mathrm{y}^T{}_n$ is the training labels which are extracted the QP value from x264 ABR rate control method. Especially, each input sequence x_L at a time step corresponding to a label y, so y^T has length T time step sequence $\langle y^1,\ y^2, \dots, y^T \rangle$ as one training label. The loss function is minimized by the stochastic gradient descent with the standard back propagation [14].

4. THE LSTM-BASED RATE CONTROL FOR X264

Fig. 2 shows the architecture of proposed LSTM-based rate control method implement in x264. When giving a P frame to x264 encoder, the pre-analysis is provided to extract the input feature sequence $\langle x_L^1,\ x_L^2, \dots, x_L^T \rangle$ with time step T. Then the QP of current encoding frame is predicted by the proposed RCLSTM network. We employ a "sliding window" mechanism for the QP prediction, which consider the temporal correlation from the encoded P frames. The details are descripted in Section 4.2. However, the proposed RCLSTM network cannot perfectly predict the accurate QP matching the target bits for each frame. The QP refinement strategy is thereby proposed to further improve the QP value based on the bitrate information. Finally, the modified QP is used for the P frame coding.

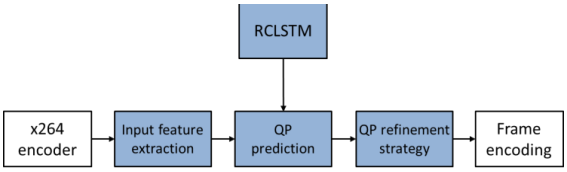


Figure. 2 The framework of the proposed rate control method

4.1 Input Feature Extraction

In x264 encoder, most of input features x_L can be extracted based on the exist function, including the MSE, SSIM, coded bits and QP extracted in last encoded frame. The SATD value can be extracted from the pre-analysis process as mentioned in section 2.2. However, the coded bit of current frame is an unknown input feature before encoding. Thus the bit allocation of current frame is proposed to replace the coded bits of current frame for RCLSTM forward prediction. The allocated bits for current frame can be expressed as:

$$P_{alloc} = \frac{R_{target} + (B_{total} - B_{wanted})}{fps} \cdot \frac{P_{coded}}{I_{coded} + P_{coded}} \quad (16)$$

where I_{coded} and P_{coded} are the total encoded bits from the previous I frame and P frames, respectively. R_{target}, B_{total}, and B_{wanted} are the same definition described in section 2.2. Actually, we expect that the bit allocation is closed to the target bits of current frame, which give a more accurate QP prediction.

4.2 QP Prediction

In feature extraction, the input features of current frame cannot present the temporal correlation with the previous encoded frames because only one input feature in current time step is considered. On the other hand, we suppose that the last output of RCLSTM achieved the most temporal correlation. Therefore, we propose a "sliding window" mechanism to predict the QP.

Firstly, the number n of previous encoded frames is selected as the input features sequence during the feature extraction. However, the n is commonly smaller than the time step T of RCLSTM at the beginning of the x264 encoding, which means the collected input feature sequences are insufficient. Thus we set a dynamic value of time step in forward processing of RCLSTM network. As shown in Fig. 3,when the number of encoded frames n smaller than the time step T of RCLSTM, the time step reset to n in the forward processing of the network. The predicted QP for current frame coding is the n-th output of RCLSTM. On the contrary, when the encoded frames are sufficient for a sliding window with time step T, we select the last (T-th) output of RCLSTM forward processing as the predicted QP for current frame coding. On the other hand, when the encoded frames are greater than the time step T, the number T of previous encoded frames is considered as the input features sequence. For example, the QP of current frame T is predicted by proposed RCLSTM with the input features sequence $\langle x_L^1,\ x_L^2, \dots, x_L^T \rangle$. For the next encoding frame T+1, the "first-input and last-output" property is applied to sliding window for input feature sequence extraction. It means that only the input feature x_L^1 is removed and the input feature x_L^{T+1} is added to consist of the new input feature sequence $\langle x_L^2,\ x_L^3, \dots, x_L^{T+1} \rangle$.

4.3 QP Refinement Strategy

In practical application, the proposed RCLSTM network cannot give a perfect coding performance in x264. Three cases are not considered: In training stage, a training sample consists of a constant input feature and a constant label from a completely encoding in x264, which means the training data cannot be updated immediately while training. Second, the training labels are extracted from ABR rate control method during the x264 encoding. Actually, the ABR method cannot be seen as a ground truth method keeping the best performance both bitrate and distortion. On the other hand, the trained RCLSTM model would generate prediction error in practical forward prediction. Therefore, the QP refinement strategy is necessary applied after the RCLSTM prediction. Firstly we keep use the $overflow[i]$ parameter as mentioned in equation (4) to refine the QP value, where i present the encoding frame number. The *overflow* closes to 1 which means that the actual encoded bitrate is closed to target

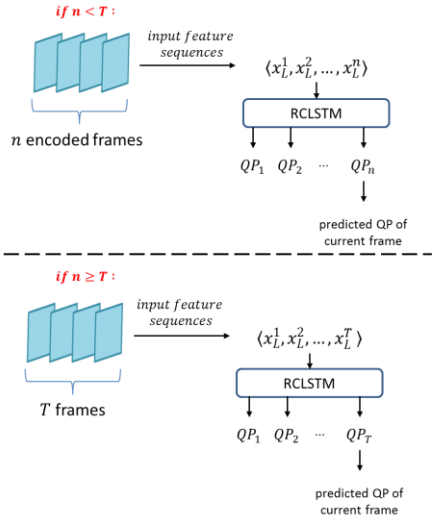


Figure. 3 Sliding window mechanism for QP prediction

QP refinement strategy

Input:

QP_{LSTM}: Predicted QP by RCLSTM forward process.

QP_{prev} : The QP_{refine} of last encoded frame.

$overflow[i]$: QP refinement parameter

Output:

Refined QP_{refine} by QP refinement strategy

Step:

1. Modify the QP_{LSTM} value by equation (5), (3), (5) sequentially.
2. **if** $|overflow[i-2]-1| > |overflow[i-1]-1|$ && $|overflow[i-1]-1| > |overflow[i]-1|$ && $|overflow[i]-1| < \xi$
   $QP_{refine} = QP_{LSTM}$,
   else if $|overflow[i-1]-1| > |overflow[i]-1| + \eta$ && $|overflow[i]-1| < 2\xi$
   $QP_{refine} = (QP_{LSTM} + QP_{prev}) / 2$
   else if $|overflow[i-1]-1| < |overflow[i]-1|$
   $QP_{refine} = overflow[i] > 1 ?$ $QP_{LSTM} + 0.5 : QP_{LSTM} - 0.5$
   end if
3. Output the QP_{refine} in range: $[QP_{prev} - 2, QP_{prev} + 2]$

bitrate. Then we use the tendency of *overflow* to refine the QP value. The following pseudo codes show the whole process of QP refinement. The parameter ξ and η are set to 0.5 and 0.002 as the threshold condition of overflow, respectively. Finally, we use the refined QP for current frame encoding.

5. EXPERIMENTAL RESULTS

5.1 Implementation

As shown in Table 2, we use the HM test sequence [15] and the surveillance video as the training data. The training samples are extracted from the x264 ABR rate control with different target bitrate. Additionally, thought the training data consists of HM sequence and surveillance video, we extract the training samples for each encoded frame to further expand the training dataset. For instance, the *RaceHorses* have 500 frames, and the time step set to 40 for RCLSTM training. Then the size of $460 \times 40 \times 6$ data are be extracted as the part of input data, where 460 training samples contain 6 input feature with 40 time step. Finally, the time step T is set to 40, and about 50,000 training samples are collected for LSTM training.

In RCLSTM training, we use caffe [16] to train the network. The initial learning rate is set to 0.0001, and the learning rate policy "step" is applied with 0.5 gamma and 500,000 stepszie. The batch size is set to 8, and we trained the RCLSTM network for 2,500,000 iterations. The NVIDIA GTX 1080Ti GPU are equipped for training. In implementation, the environment system is Ubuntu-14.04.1 with Intel(R) Xeon(R) CPU E5-2650 v4 at 2.20GHz and 6GB RAM memory. We test the proposed RCLSTM rate control method in x264 encoder without GPU.

Table 2. The sequences for RCLSTM training

Size	Training sequence	Target bitrate
2560 × 1600	*PeopleOnStreet*、*Traffic*	2 Mbps, 2.5 Mbps, 3 Mbps, 3.5 Mbps, 4 Mbps, 4.5 Mbps, 5 Mbps, 5.5 Mbps
1920 × 1080	*BasketballDrive*、*BQTerrace*、*Cactus*、*cheliangdaolu1*、*cheliangdaolu2*、*Talking*、*Walking*	1 Mbps, 1.5 Mbps, 2 Mbps, 2.5 Mbps, 3 Mbps, 3.5 Mbps, 4 Mbps, 4.5 Mbps
1280 × 720	*vidyo1,vidyo3,vidyo4,atm_people_coming,day_cross_car_coming,day_cross_cars_waiting_signalevening_cross,evening_cross_small_motion,garden_complex_media_motion,night_cross_bus_people*	1 Mbps, 1.5 Mbps, 2 Mbps, 2.5 Mbps, 3 Mbps, 3.5 Mbps, 4 Mbps, 4.5 Mbps
832 × 480	*BasketballDrill*、*BQMall*、*PartyScene*、*RaceHorses*	1 Mbps, 1.5 Mbps, 2 Mbps, 2.5Mbps, 3 Mbps, 3.5 Mbps, 4 Mbps, 4.5 Mbps
416 × 240	*BasketballPass*、*BlowingBubbles*、*BQSquare*、*RaceHorses*	0.5 Mbps, 1 Mbps, 1.5 Mbps, 2 Mbps, 2.5 Mbps, 3 Mbps, 3.5 Mbps, 4 Mbps

5.2 Results and Analysis

In experimental results, we use two metrics to evaluate the performance of proposed rate control method. The Bjøntegaard delta bit rate (BD-rate) [17] is utilized to evaluate the coding performance of the proposed method. The target bit matching rate is given as:

$$BRAC = (1 - \frac{|R_{target} - R_{coded}|}{R_{target}}) \times 100\% \quad (17)$$

where R_{coded} is the coded bit rate from rate control. For result evaluation, the configuration of x264 encoder is shown in Table 3. In addition, we disable the adaptive quantization, buffer occupancy and scene change in x264, which would influence the comparison between the proposed method and x264 ABR rate control at frame level. Only the first frame is encoded as I frame, then all the P frames are encoded continuously.

Table 3. The coding configuration of x264

Profile	main
Encoding speed	medium
Rate control method	ABR
Adaptive quantization	off
Buffer occupancy	off
Scene change	off
I frame interval	-1
Target bitrate (Mbps)	1,2,3,4

Table 4. The comparison results of proposed method

HM Test Sequence	Proposed method		x264 ABR
	BD-rate	BRAC	BRAC
ClassB_Tennis_1920x1080	1.7%	98.60%	90.42%
ClassB_Kimono_1920x1080	2.1%	99.55%	85.28%
ClassB_ParkScene_1920x1080	−0.4%	99.19%	96.51%
ClassC_Flowervase_832x480	−0.6%	98.57%	88.42%
ClassC_Keiba_832x480	0.7%	98.83%	96.59%
ClassD_Flowervase_416x240	0.8%	97.26%	87.19%
ClassD_Keiba_416x240	0.5%	98.93%	97.13%
ClassE_FourPeople_1280x720	−4.3%	99.99%	91.83%
ClassE_Johnny_1280x720	−5.7%	98.36%	98.08%
ClassE_KristenAndSara_1280x720	−6.3%	99.76%	98.20%
Average	**−1.2%**	**98.90%**	**92.97%**

Table 5. The property of surveillance video

Surveillance Test Sequence	Envir.	Light.	Motion
yuxuedaolu	Outdoor	Day	High
bus	Outdoor	Day	Medium
Crew	Indoor	Day	Medium
darkroom1	Indoor	Night	High
darkstreet1	Outdoor	Night	High
roads_coming_leaving	Outdoor	Day	Medium
indoor_complex_media_motion	Indoor	Day	Medium
day_cross_midday_complex	Outdoor	Day	Medium
subway_higher_very_complex	Indoor	Day	High
corridor_people_coming_complex	Indoor	Night	High

Table 6. The comparison results of proposed method

Surveillance Test Sequence	Proposed method		x264 ABR
	BD-rate	BRAC	BRAC
yuxuedaolu_1920x1080	−1.7%	99.44%	98.86%
bus_1280x720	−1.7%	99.76%	98.90%
Crew_1280x720	1.5%	98.93%	92.93%
darkroom1_1280x720	−0.1%	99.81%	97.87%
darkstreet1_1280x720	0.5%	99.82%	97.22%
roads_coming_leaving_1280x720	−3.8%	99.90%	95.15%
indoor_complex_media_motion_ 1280x720	−1.9%	99.94%	96.89%
day_cross_midday_complex_ 1280x720	−2.3%	99.66%	96.78%
subway_higher_very_complex_ 1280x720	0.1%	99.75%	96.79%
corridor_people_coming_complex_ 1280x720	−0.5%	99.96%	99.41%
Average	**−1.0%**	**99.70%**	**97.08%**

Test for HM sequence. We compare the BD-rate and target bit matching rate to x264 ABR method. Each test sequence is tested by ABR method with 4 different target bitrate shown in Table 3. Especially, the sequence with 416×240 is tested in 0.5, 1.5, 2.5, 3.5 target bitrate, respectively. As shown in Table 4, we can see that the BD-rates are achieved up to − 6.3% at sequence *KristenAndSara*. Meanwhile, the target bit matching rate is 98.90% on average which is also outperforms to ABR method with 92.97% target bit matching rate.

Test for surveillance video. Table 5 represents the property of the surveillance video. The coding performance result is shown in Table 6. It can be concluded that the proposed method is also adopted in surveillance video with various video content.

6. CONCLUSION

In this paper, a novel LSTM-based rate control method is proposed for x264. The temporal correlation is founded and a more accurate QP is predicted by the LSTM network. Then the QP refinement strategy improves the value which is closed to the target bits for current frame coding. In future work, we expect to apply the proposed method to other encoder, i.e. HM, vp9, or even integrated in chip.

7. REFERENCES

[1] T. Wiegand, G.J. Sullivan, G. Bjøntegaard, A. Luthra. 2003. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 13, n. 7, pp. 560–576. Aug. 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand. 2012. "Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 22, n. 12, p. 1649–1668, Dec. 2012.

[3] Z. Li, F. Pan, K. P. Lim, G. Feng, X. Lin and S. Rahardja. 2003. Adaptive basic unit layer rate control for JVT. Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G012, Pattaya II, Thailand, Mar. 2003.

[4] H. Choi, J. Nam, J. Yoo, D. Sim, and I. V. Bajić. 2012. Rate control based on unified RQ model for HEVC. JCTVC-H0213, 8-th JCTVC meeting, San Jose, CA, USA, Feb. 2012.

[5] H. Choi, J. Yoo, J. Nam, D. Sim, and I. V. Bajić. 2013. Pixel-wise unified rate-quantization model for multi-level rate control. *IEEE J. Sel. Topics Signal Process.*, v. 7, n. 6, p. 1112–1123, Dec. 2013.

[6] B. Li, H. Li, L. Li, and J. Zhang. 2012. *Rate Control by R-Lambda Model for HEVC*. In JCTVC-K0103, 11th Meeting, Shanghai, China, Oct. 2012.

[7] B. Li, H. Li, L. Li, and J. Zhang. 2014. λ domain rate control algorithm for high efficiency video coding. *IEEE Transactions on Image Process.*, v. 23, n. 9, p. 3841–3854, Sep. 2014.

[8] x264 encoder [Online] Available: http://www.videolan.org/developers/x264.html.

[9] Y. Dai, D. Liu, and F. Wu. 2017. A Convolutional Neural Network Approach for Post-Processing in HEVC Intra Coding. *International Conference on Multimedia Modeling*. Springer, Cham, p. 28-39, 2017.

[10] M. Xu, T. Li, Z. Wang, X. Deng, R. Yang, and Z. Guan. 2018. Reducing complexity of HEVC: a deep learning approach. *IEEE Transactions on Image Process.*, v. 27, n. 10, p. 5044–5059, June 2018.

[11] W. Gao, S. Kwong, and Y. Jia. 2017. Joint machine learning and game theory for rate control in high efficiency video coding. *IEEE Transactions on Image Process.*, v. 26, n. 12, p. 6074–6089, Dec. 2017.

[12] Y. Li, B. Li, D. Liu, and Z. Chen. 2017. A convolutional neural network-based approach to rate control in HEVC intra coding. *Proceedings of IEEE Visual Communications and Image Processing* (VCIP). Dec. 2017

[13] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of Computer Vision and Pattern Recognition*, p. 2625-2634, 2015

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278-2324, 1998.

[15] F. Bossen. 2012. Common test conditions and software reference configurations," San Jose, USA, JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCT-VC H1100 (m24011), Feb. 2012.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In ACM MM, 2014.

[17] G. Bjøntegaard. 2001. Calculation of average PSNR differences between RD-curves. Technical Report VCEG-M33, ITU-T SG16/Q6, Austin, TX, USA.

Improvement on Demosaicking in Plenoptic Cameras by Use of Masking Information

Hyunji Cho
SangMyung University
South Korea
chohj0510@naver.com

Joungeun Bae
SangMyung University
South Korea
wjddms5810@naver.com

Hyeonah Jung
SangMyung University
South Korea
201411226@sangmyung.kr

Eunju Oh
SangMyung University
South Korea
essenti10@naver.com

Hoon Yoo
SangMyung University
South Korea
hunie@smu.ac.kr

ABSTRACT

Most cameras use a color filter array (CFA) to capture color images. This array has only one color among red, green and blue per pixel and so, each pixel on the sensor lacks two color channels which can be reconstructed by a process called demosaicking. This paper proposes a demosaicking method for a plenoptic camera. Plenoptic raw images have a particular lenslet structure which must be formed the empty space, unlike conventional images. These vacant spaces exist because the lens is circular. The existing demosaicking methods degrade the image quality because it interpolates missing colors, including information on the empty space. Recently, a method using information about empty spaces has been studied. However, this method uses empty space information only for G-plane. To improve this method we applied information about empty space to R/B-plane too. Experimental results show that the performance of the proposed method is superior. As compared with the existing demosaicking algorithms, we show that our solution produces the best average demosaicking performance both objectively and subjectively.

CCS Concepts

Theory of computation~Backtracking

Keywords

Demosaicking; color demosaicking; interpolation; Plenoptic camera; Light field camera; Light field interpolation; Plenoptic interpolation

1. INTRODUCTION

The plenoptic camera provides both spatial and angle information of a scene, unlike the existing camera which provides only 2D spatial coordinates. By placing a microlens array between the main lens and the sensor, the plenoptic camera obtains the intensity and color as well as the direction of the light bundles that enter the camera. Captured data provides a matrix of horizontally and vertically aligned views from slightly different points of view over the scene. These images can be post-processed for either synthesizing depth estimation or for post-capture refocusing [1-2].

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301530

The angular resolution of the plenoptic cameras, in essence, has a lower spatial resolution of images. Ultra-high resolution sensors are commonly used to overcome the spatio-angular tradeoff. However, the resolution of the resulting image is still low. Therefore, a method for improving the image quality is needed. Demosaicking is one of those methods.

Demosaicking is a technique that predicts and interpolates a missing color information according to the CFA type when creating a color image using a single sensor [3-4]. In general, demosaicking requires minimizing noise such as aliasing, chromatic aberration, zippering, etc., which affect image quality. So, many algorithms about demosaicking have been studied [5-11].

Plenoptic raw images have a different structure than existing images. These images consist of a set of elemental images through a microlens array. Recently, various demosaicking techniques using the structural feature of plenoptic images have been studied [12-15]. The lastest method has been proposed demosaicking about the empty space caused by a circular lens array, which is the most prominent feature of the plenoptic image [12-13]. These demosaicking methods apply the Hamilton-Adams method to the information about the empty space between lenses.

This paper has improved demosaicking [12]. Although demosaicking [12] uses mask information only for G-plane, the proposed method applies mask data to R/B-plane to omit unnecessary processes about empty space and improve performance. The proposed method shows objective and subjective performance improvement through experiments.

2. RELATED WORK

2.1 Hamilton-Adams Method (HA)

HA is a low-complexity interpolation method considering the correlation and directionality between planes. Color information of other planes is used together to detect high frequency when judging the directionality. Then, a low-frequency component of the same plane and a high-frequency component of another plane are combined by taking advantage of the high correlation between planes, when predicting color information. Since the interpolated color information has both a low-frequency component and a high-frequency component, the predicted value is relatively accurate. Because the HA method is based on the Bayer pattern, the difference between the amount of G color information and the amount of R/B color information is remarkable. Therefore, HA interpolates R/B-plane considering the reconstructed G-plane after interpolating G-plane with a lot of information first.

(a) **(b)**

Figure 1. Pixels in use for reconstruction in HA method (a) pixels for reconstruction of G-plane (b) pixels for reconstruction of R/B-plane

2.1.1 G-plane Interpolation

G-plane has a quincunx shape so, there are pixels in the horizontal and vertical directions but no pixels in the diagonal direction. Therefore, demosaicking is performed in both horizontal and vertical directions. First, we obtain the gradient in the vertical and horizontal directions and then calculate the interpolation value along each direction. The calculated value is determined for a direction with a small gradient value. The pixel used for demosaicking is as shown in Fig. 1 (a). For example, when interpolating x_5 position, Eq. (1) means interpolated in the horizontal direction, and Eq. (2) means interpolated in the vertical direction.

$$x_5\mathrm{H} = (x_4 + x_6)/2 + (-x_3 + 2x_5 - x_7)/4 \quad (1)$$

$$x_5\mathrm{V} = (x_2 + x_8)/2 + (-x_1 + 2x_5 - x_9)/4 \quad (2)$$

2.1.2 R/B-plane Interpolation

The R/B-plane has only one pixel value in a 2x2 block. Thus, interpolation should be performed for the remaining three positions. Reconstruction of R/B-plane is performed differently depending on the position of the missing pixel. The pixels used for R/B-plane interpolation are shown in Fig. 1 (b), and both R-plane and B-plane has the same shape. First, the x_2 and x_4 positions where the same plane pixel exists on the left and right or upper and lower sides are interpolated using the low-frequency component of the same plane and the high-frequency component of the G-plane. Eq. (3) means that there are components of the same plane on the left and right and Eq. (4) means that there are components of the same plane on the upper and lower sides.

$$x_2 = (x_1 + x_3)/2 + (-x_1 + 2x_2 - x_3)/2 \quad (3)$$

$$x_4 = (x_1 + x_7)/2 + (-x_1 + 2x_4 - x_7)/2 \quad (4)$$

Finally, the x_5 position where the same planar pixel exists on diagonal and off-diagonal lines interpolates by setting the edges to diagonal or off-diagonal directions. It calculates a gradient diagonally or off-diagonally in the same order as the G-plane interpolation method, and interpolation values along each direction are calculated. The interpolation values according to the selected direction are calculated by the Eq. (5) and Eq. (6). Eq. (5) means that there is a component of the same plane in the diagonal direction, and Eq. (6) means that there is a component of the same plane in the diagonal direction.

$$x_5\mathrm{N} = (x_1 + x_9)/2 + (-x_1 + 2x_5 - x_9)/2 \quad (5)$$

$$x_5\mathrm{P} = (x_3 + x_7)/2 + (-x_3 + 2x_5 - x_7)/2 \quad (6)$$

2.2 Masking Based Method

Mask data based demosaicking is a method that utilizes information about the empty space between the lenses. Empty space is unnecessary information because it does not have any information about the image. Therefore, it eliminates unnecessary arithmetic operations and improves accuracy.

In mask data based demosaicking, mask data is applied only to G-plane. The R/B-plane is demosaicked by the HA method. The pixel used in G-plane reconstruction is shown in Fig. 1 (a). G-plane interpolation is divided into three methods. First, if the pixel to be restored is not included in the mask area, it means an empty space. So, in this case operation is unnecessary. Second, the pixel to be restored is in the mask area, and all surrounding pixels needed for restoration are in the mask area. Since there is no unnecessary information at this time, the HA method is applied. Finally, the pixel to be restored is in the mask area, but only a few neighborhood pixels are present in the mask area. In this case, only the pixels in the mask area are used for interpolation. For example, the method of interpolating the x_5 position is shown in Eq. (7). Mask data is denoted by M, where the value of a pixel in the mask is 1 and the value of a pixel in non-mask is 0.

$$\widehat{x_5} = \left\{\left(\sum_{k=1}^{4} M_{2k} \cdot x_{2k}\right) \Big/ \sum_{k=1}^{4} M_{2k}\right\} + \left\{\left(\sum_{k=1}^{4} M_{2k-1} \times x_5 - \sum_{k=1}^{4} M_{2k-1} \cdot x_{2k-1}\right) \Big/ \left(2 \times \sum_{k=1}^{4} M_{2k-1}\right)\right\} \quad (7)$$

3. PROPOSED SYSTEM

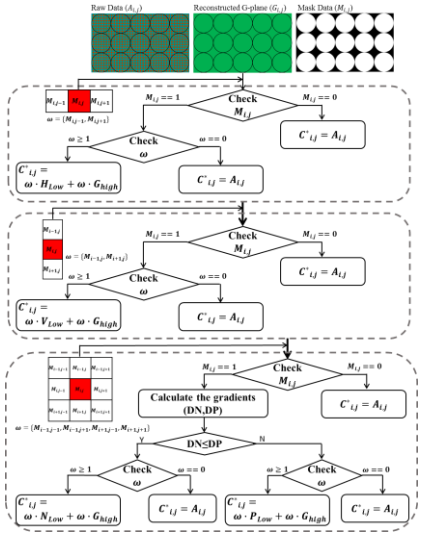


Figure 2. Block Diagram of Proposed Method

This paper proposes a demosaicking method by applying mask data to R/B-plane by improving mask data based method. Mask data is extracted from the plenoptic raw image. Plenoptic cameras acquire images through an array of circular lenses, resulting in empty space in the image. The empty space has no information about the image. Thus, it is used as mask data to omit image processing for empty space. This eliminates unnecessary arithmetic operations and improves the accuracy of the arithmetic result.

The proposed method is based on the Bayer pattern. Therefore, since the G components are twice the R/B components, the G-plane is first interpolated and then the R/B-plane is interpolated. G-plane applies the masking based method. The block diagram for the R/B-plane interpolation of the proposed method is shown in Fig. 2. There is only one pixel in the R or B color in the 2x2 block.

So, interpolation for the remaining three pixels is required. The missing color component is reconstructed using the low-frequency of the same color as the color of the pixel to be restored and the high-frequency of the G component. Used pixels must be in the mask. Interpolation is performed in the unit of the mask because the mask data is utilized. The interpolation method is divided into three cases for three pixels. However, all cases are executed when the pixel to be restored is in the mask, and interpolation is not executed when it does not exist in the mask since it means that it is an empty space.

In the first case, pixels of the same color exist on the left and right sides. In this case, the interpolation method depends on whether the pixels used for reconstruction is in the mask or not, when the pixel to be reconstructed exists in the mask. For example, Eq. (8) is interpolation of x_2 from Fig. 1. (b).

$$\widehat{x_2} = \begin{cases} x_2 & if \sum_{k=1}^{2} M_{2k-1} = 0 \\ \left\{\left(\sum_{k=1}^{2} M_{2k-1} \cdot x_{2k-1}\right) / \sum_{k=1}^{2} M_{2k-1}\right\} + & \\ \left\{\left(\sum_{k=1}^{2} M_{2k-1} \times G_2 - \sum_{k=1}^{2} M_{2k-1} \cdot G_{2k-1}\right) / \sum_{k=1}^{2} M_{2k-1}\right\} & else \end{cases} \quad (8)$$

The second case is that pixels of the same color are at the upper and lower sides of the pixel to be interpolated. At this time, if the neighbor pixels used for reconstruction are out of the mask area, the color is not restored. For example, when interpolating x4, Eq. (9) is obtained.

$$\widehat{x_4} = \begin{cases} x_4 & if \sum_{k=1}^{2} M_{6k-5} = 0 \\ \left\{\left(\sum_{k=1}^{2} M_{6k-5} \cdot x_{6k-5}\right) / \sum_{k=1}^{2} M_{6k-5}\right\} + & \\ \left\{\left(\sum_{k=1}^{2} M_{6k-5} \times G_4 - \sum_{k=1}^{2} M_{6k-5} \cdot G_{6k-5}\right) / \sum_{k=1}^{2} M_{6k-5}\right\} & else \end{cases} \quad (9)$$

Finally, the four pixels in the diagonal direction of the pixel to be restored are the same color. In that case, since there are two cases of diagonal direction and off-diagonal direction, it is necessary to interpolate it in consideration of directionality. The missing pixel is interpolated in the corresponding direction by obtaining the gradient for each direction. For example, when interpolating x_5, Eq. (10) is an interpolated expression when the gradient is diagonal and Eq. (11) is an interpolated expression when the gradient is off-diagonal. If the surrounding pixels are not in the mask area, do not interpolate.

$$\widehat{x_5} = \begin{cases} x_5 & if \sum_{k=1}^{2} M_{4k-1} = 0 \\ \left\{\left(\sum_{k=1}^{2} M_{4k-1} \cdot x_{4k-1}\right) / \sum_{k=1}^{2} M_{4k-1}\right\} + & \\ \left\{\left(\sum_{k=1}^{2} M_{4k-1} \times G_5 - \sum_{k=1}^{2} M_{4k-1} \cdot G_{4k-1}\right) / \sum_{k=1}^{2} M_{4k-1}\right\} & else \end{cases} \quad (10)$$

$$\widehat{x_5} = \begin{cases} x_5 & if \sum_{k=1}^{2} M_{8k-7} = 0 \\ \left\{\left(\sum_{k=1}^{2} M_{8k-7} \cdot x_{8k-7}\right) / \sum_{k=1}^{2} M_{8k-7}\right\} + & \\ \left\{\left(\sum_{k=1}^{2} M_{8k-7} \times G_5 - \sum_{k=1}^{2} M_{8k-7} \cdot G_{8k-7}\right) / \sum_{k=1}^{2} M_{8k-7}\right\} & else \end{cases} \quad (11)$$

4. SIMULATION

A comparison with the existing method was carried out to evaluate the performance of the proposed method subjective and objectively. The experiment revealed that results of conventional methods are different according to the background color, which is an empty space. Experimental results are better than when the background color is white when the background color is black because the experimental image is entirely dark. In this paper, the background color is gray. The experimental images were taken from 4 raw plenoptic images of size 7560 x 5250 with the Lytro Illum camera, as specified in Fig. 3. We measured the color signal-to-noise ratio (CPSNR) of the interpolated images of both linear interpolation (LI), HA [6], masking based method [12] and proposed demosaicking with respect to the original image.

Figure 3. Set of Testing Images

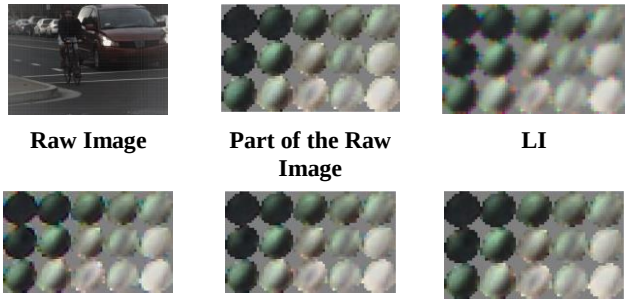


Figure 4. Results of Experiment

Table 1. CPSNR Results of Different Demosaicking Methods

	LI	HA	Mask Based Method	Proposed Method
Bike	25.73	30.69	40.80	43.55
Building	22.18	27.69	40.25	41.16
Flower	20.67	28.56	41.93	42.50
People	22.85	28.13	40.57	41.47
AVE	22.86	28.77	40.89	**42.17**

Table 1 tabulates the CPSNR performance of various algorithms. The proposed method results in higher overall CPSNR values than LI et al. Fig. 4 shows a partial enlargement of the experimental results. In Fig. 4, we can see that there is no zippering phenomenon near the lens edge and color interpolation is also good. Therefore, it can be seen that the proposed method is superior in objectivity and subjective quality.

5. CONCLUSION

In this paper, we propose a method to improve masking based demosaicking. Plenoptic cameras acquire images through a microlens array. Since the lens has a circular shape, the plenoptic raw image has an empty space between the lenses. Empty space interferes with image processing because there is no information. Therefore, the information about empty space is used as mask data to reduce the computation and improve the performance. Experiments show that the proposed method is superior to other demosaicking algorithms objectively and subjectively. These results are expected to improve imaging quality by expanding the demosaicking technique of plenoptic images.

6. ACKNOWLEDGMENTS

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.2017-0-00515, Development of integraphy content generation technique for N-dimensional barcode application)

7. REFERENCES

[1] X. Xiao, B. Javidi, M. Martinez-Corral, and A. Stern, ‘‘Advances in three-dimensional integral imaging: Sensing, display, applications,’’ Appl. Opt., vol. 52, no. 4, pp. 546–560, 2013.

[2] Y. Lee and H. Yoo, "Three-dimensional visualization of objects in scattering medium using integral imaging and spectral analysis," Elsevier Optics and Lasers in Engineering, vol. 77, no. 2, pp. 31-38, Feb. 2016.

[3] R. Lukac, Single-sensor imaging: methods and applications for digital cameras, CRC Press, 2009.

[4] O. Losson, L. Macaire, Y. Yang, Advances in Imaging and Electron Physics, Elsevier, vol. 162 pp. 173-265, 2010.

[5] H. Yoo, D. Yoon, and D. Kim “Decimation-interpolation structures for image communications and improvement using the lifting scheme,” IEEE Trans. Consumer Electronics, vol. 56, no. 4, pp. 2669-2677, Nov. 2010.

[6] J. E. Adams and J. F. Hamilton Jr., "Adaptive color plane interpolation in single color electronic camera," US patent, 5506619, 1996.

[7] J. S. Jimmy Li and S. Randhawa, "Color filter array demosaicking using high-order interpolation techniques with a weighted median filter for sharp color edge preservation," IEEE Trans. Image Process., vol. 18, no. 9, pp. 1946-1957, Sep. 2009.

[8] R. Lukac, K. N. Plataniotis, D. Hatzinakos and M. Aleksic, "A novel cost effective demosaicing approach," IEEE Transactions on Consumer Electronics, vol. 50, no. 1, pp.256-261 2004.

[9] K.-H. Chung and Y.-H. Chan, "Color demosaicing using variance of color differences," IEEE Trans. Image Process., vol. 15, no. 10, pp.2944-2955 2006.

[10] A. Buades, B. Coll, J.-M. Morel, and C. Sbert, “Self similarity driven demosaicking,” IEEE Transaction on Image Processing, vol. 18, no. 6, pp.1192-1202, 2009.

[11] E. Dubois, “Frequency-domain methods for demosaicking of bayer-sampled color images,” IEEE Signal Processing Letters, vol. 12, pp. 847-850, 2005.

[12] Cho, Hyunji, and Hoon Yoo. "Masking Based Demosaicking for Image Enhancement Using Plenoptic Camera." International Journal of Applied Engineering Research vol. 13, no.1, pp. 273-276, 2018.

[13] P. David, M. Le Pendu, & C. Guillemot, “White lenslet image guided demosaicing for plenoptic cameras,” In MMSP 2017-IEEE 19th International Workshop on Multimedia Signal Processing. 2017.

[14] Z. Yu, J. Yu, A. Lumsdaine, and T. Georgiev. “An analysis of color demosaicing in plenoptic cameras,” In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on, pp. 901–908, IEEE, 2012.

[15] T. Lian, and C. Kyle, "Demosaicing and Denoising on Simulated Light Field Images,” 2016.

[16] M. Seifi, N. Sabater, V. Drazic, and P. Perez, “Disparity guided demosaicking of light field images,” in IEEE International Conference on Image Processing (ICIP), 2014.

MLN: Moment localization Network and Samples Selection for Moment Retrieval

Bo Huang
Cooperative Medianet Innovation Center,
Shanghai Jiao Tong University, China
bo_huang@sjtu.edu.cn

Ya Zhang
Cooperative Medianet Innovation Center,
Shanghai Jiao Tong University, China
ya_zhang@sjtu.edu.cn

Kai Yu
Cooperative Medianet Innovation Center,
Shanghai Jiao Tong University, China
kai.yu@cs.sjtu.edu.cn

ABSTRACT
Moment retrieval is a hot problem recently. Given a video, people want to retrieve the clip that matches some semantic meaning. This problem is difficult because both video understanding and language understanding are needed and as a problem of cross-modality retrieval, cross-modality method and similarity metric design are important. Previous research established a general framework for moment retrieval. In this paper, we refine the framework and name it moment localization network and propose two novel sample selection methods to improve training of the model. We do experiments on two large datasets: TACoS and DiDeMo. Results show we outperform previous state-of-the-art method and our sample selection method makes improvement.

CCS CONCEPTS
• Computing methodologies → Matching

Keywords
moment retrieval, samples selection, video

1. INTRODUCTION
Video retrieval by natural language has long been studied. The task is challenging because both video understanding and natural language understanding are required, and to calculate similarity, a cross-modal method is needed. Traditional video retrieval problem retrieves a video from a trimmed video collections therefore the usage is limited. In real world, most videos are untrimmed and long, and people want to retrieve moments among the long video. This is moment retrieval.

Natural language video retrieval methods aim to retrieve a specific video given a natural language query.[11][20]incorporate deep video-language embedding similar to image-language embedding proposed by[3][17][10] localize natural language phrases spatially in images. A popular model framework of video retrieval is video feature extraction, language feature extraction, cross-modal similarity calculation. Traditional method use rigid math function to calculate similarity. For example,[16] used Canonical correlation analysis(cca) to calculate the similarity. As deep neural network(dnn) showed its effectiveness, dnn was introduced in similarity calculation. [18]uses fully connected layers(fc) to project both video and language representations into a common space and use fc to learn to the metric to calculate similarity. [5] projects the same and directly uses minus euclidean distance as the similarity score in the common space. [4] concatenate several cross-modal features and shows improvement in video retrieval recall. [8][2][1][15] consider aligning textual instructions to videos.

Our contributions are as follow: (1)We introduce a novel network to solve this problem. (2) We introduce data selection as a novel promotion method for moment retrieval. (3) We do experiments on two large dataset and outperform previous state-of-the-art method.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301538

2. MOMENT LOCALIZATION NETWORK
MLN includes two parts: a novel network to match video cliplanguage query and propose matching area of video clip and a novel way to do data augmentation and selection in model training.

2.1 Problem Formulation
We denote a video as $V = \{ f_i \}_{i=1}^{N}$, N is the frame number of the video. A video is associated with several temporal sentence annotations: $A = \{(s_j, ts_j, te_j)\}_{j=1}^{M}$, s_j is a natural language sentence, the sentence matches a part of the video. ts_j and te_j denotes the start frame and end frame of the matched part. M is the number of annotations. Training data are the video and their sentence annotations. The task is to predict ts_j, te_j by given V and s_j.

2.2 Match and Propose Network
In this section, we describe our match and propose network(MPN). As illustrated in Fig 1, MPN contains 4 parts: a video encoder to extract feature for video clips, a sentence encoder to extract feature for sentence query, a cross-modal module to map features in different modality into a common space and a match and propose module to compute match score and propose refined video clip.

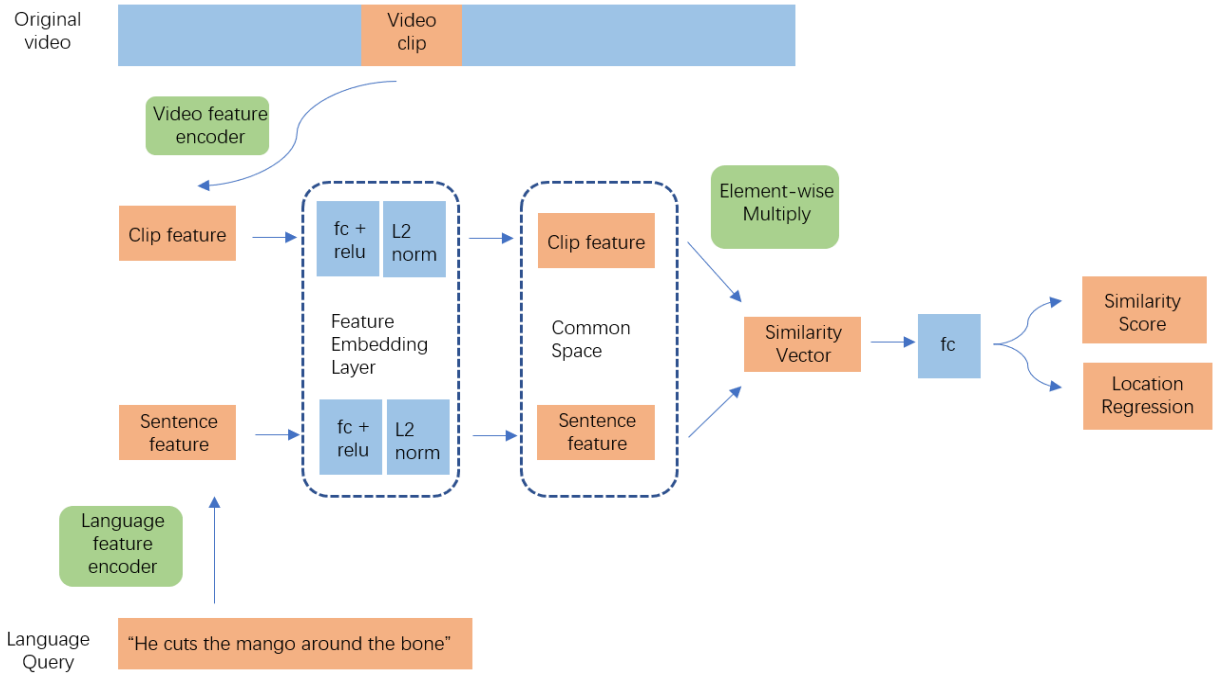


Figure 1. Moment Localization Network(MLN) model.

2.2.1 Visual Encoder.

For a long untrimmed video V , we generate a set of video clips $C = \{c_i\}_{i=1}^{H}, c_i = (V, ts_i, te_i)$ by temporal sliding windows, where ts_i and te_i denote the clip start frame and end frame of V , H is the number of video clips. Visual encoder is defined as a function $F_v e(c_i)$ that projects a certain clip c_i into a feature vector $f_{c_i}^{v}$ of dimension d_v . We choose C3D and TSN pretrained on image classification and fintuned on action recognition as our visual encoder. We think those models are able to captures both static and dynamic information in videos.

2.2.2 Sentence Encoder.

Sentence encoder is defined as $F_s e(s_j)$ that projects a natural language sentence into a feature vector of dimension d_s . We try training LSTM from scratch on our dataset and using pretrained skip-thought models.

2.2.3 Cross-modal Module.

Feature vectors of video $f_{c_i}^{v}$ and sentence $f_{s_j}^{s}$ are in different modality and can not be computed together directly. We use a fully connected layer and a l2 normalization layer to project $f_{c_i}^{v}$ and $f_{s_j}^{s}$ into a video-language common space as $f_{c_i}^{c}$ and $f_{s_j}^{c}$, both are at dimension d_c .

2.2.4 Match and Propose Module.

Getting two feature vectors at a common space, a useful similarity metric is inner product. Besides, we consider feature at different channel should have different importance in calculating final similarity. Therefore, we add another fully connected layer after element-multiply layer, to set each channel a learned weight. We believe a location refinement is also able to be calculated at the same way, so we add a location refinement calculation just as similarity calculation. Our similarity function can be illustrated as:

$$sim_{w-cos}(c_i, s_j) = \frac{\sum_{k=1}^{d_e} w_k * f_k^{cci} * f_k^{csj}}{\sqrt{\sum_{k=1}^{d_e} f_k^{cci^2}} * \sqrt{\sum_{k=1}^{d_e} f_k^{csj^2}}} = \frac{W f^{cci^T} f^{csi}}{||f^{cci}|| * ||f^{csi}||} \quad (1)$$

where W is a diagonal matrix of elements $\{w_1, w_2, ...w_k\}$. reg_s and reg_e are calculated same as sim. And location after refinement as:

$$loc_{r-s} = ts + reg_s$$
$$loc_{r-e} = te - reg_e \quad (2)$$

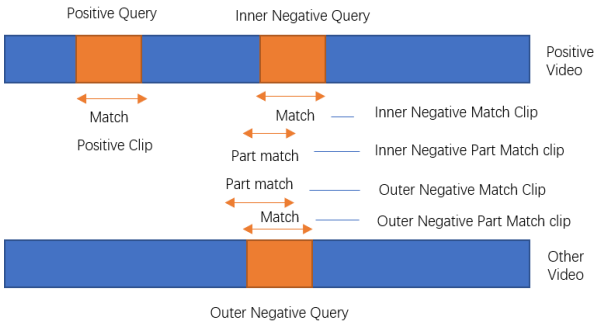


Figure 2. Samples selection methods. Basic selection method uses negative match clips. Augmentation method uses all negative clips. Effective selection uses inner negative clips

3. TRAINING AND SAMPLES SELECTION

3.1 Loss Function

At test time, by given a language query, we need to propose a clip candidate list. A rank loss is appropriate to train similarity. Our data can only inform us of matching or not. As a simplification of rank loss, we use triplet loss, which is defined as: (1) Given a language query s_j (2) Find a video clip c_i^v that matches s_j, denoted as (c_i^v, s_j), a positive pair. (3) Find a video clipv c_k^v ,that does not match s_j, denoted as (c_k^v, s_j), a negative pair. (4) Target is set to let similarity of positive clip sim_{ij} larger than similarity of any negative clip sim_{kj}. Thus similarity loss is noted as :

$$l_{sim} = max(sim_{neg} - sim_{pos} + margin, 0.0) \quad (3)$$

Where margin is proved useful in SVM. We want similarity of each pair to be in small scale and zero-centered thus we add a regularization loss for similarity.

$$l_{reg-sim} = ||sim||^2 \quad (4)$$

For simplicity, we use sliding windows to generate video proposals.

Proposals are at fixed length in this style. To generate flexible clips, we add a location refinement target. Target loss is denoted as: q

$$\begin{aligned} l_{loc_s} &= |ts_i + loc_{r-s} - ts_j^q| \\ l_{loc_e} &= |te_i - loc_{r-e} - te_j^q| \\ l_{loc} &= l_{loc_s} + l_{loc_e} \end{aligned} \quad (5)$$

Our total loss function is denoted as:

$$L = \sum_{j=1}^{n} l_{sim} + l_{reg-sim} + l_{loc} \quad (6)$$

3.2 Training Samples Selection-Basic

As procedure described in section 3.1, we define a clip c_i^v match a query s_j when iou between (ts_i, te_i) and (ts_j, te_j) larger than iou_{match} and niol for clip to query smaller than $niol_{match}$. A pair (c_i^v, s_j) is positive when c_i^v matches s_j. And we can select any unmatching pair as negative sample. A good negative sample should contain meaningful scenes instead of containing background only. Clips that matches with any query must be meaningful. Therefore, we choose negative samples (c_k^v, s_j) such that c_k^v does not match s_j and c_k^v matches any query s_l.

3.3 Training Samples Selection-Augmentation

In section 3.2, we describe valuable sample selection. As annotations are quite limited, a data augmentation method will offer lot help. Negative samples are meaningful when it matches with any query and we think clips that partly match with any query are also meaningful. We define a clip c_k^v partly match a query s_j when niol for clip to query smaller than $niol_{p-match}$.

3.4 Effective Samples Selection

In section 3.3, we describe samples augmentation. After that, negative samples are strongly augmented. However, not all samples can be same useful in introducing knowledge. Many samples introduce same knowledge and continuous training brings overfitting. A common measure of samples' efficiency is training loss, however training loss can only be acquired after feeding samples to a trained model. This will cost too much time. We find samples where positive clip is similar to negative clip have larger loss. We measure similarity of clips by cosine of their feature. The Pearson correlation coefficient between similarity and their loss is 0.286. Therefore we can select samples that are similar in clips as indicative samples. However, as there are too many available samples, time and space to calculate similarity are not acceptable. We find clips in same videos are more similar than random selection. Mean and std of cosine for clips in same video are 0.74 and 0.07, and they are 0.85 and 0.07 for random selection. Therefore we can select positive and negative clips are from a same video to make up an indicative samples. This way takes few extra time and space. Selection methods are illustrated in 2.

4. EXPERIMENTS

In this section, we describe the evaluation settings and discuss the experiment results.

4.1 Datasets

4.1.1 TACoS. .

The first dataset is constructed by [13]. This dataset was built on the top of MPII-Compositive dataset [14] and contains 127 videos. In total, there are 17344 pairs of sentence and video clips. We split it in 50% for training, 25% for validation and 25% for test. In paper [4], each training video is sampled by multi-scale temporal sliding windows with size of [64, 128, 256, 512] frames and 80% overlap. As for the testing samples, they are coarsely sampled using sliding windows with size of [128, 256] frames. They utilized C3D as the moment-level visual encoder and Skip-thoughts as the query description embedding extractor.

4.1.2 DiDeMo. .

The second dataset is constructed by [5] for languagebased moment retrieval, named the Distinct Describable Moments (DiDeMo) dataset. It includes 10,464 personal videos with duration of 25-30 seconds, 26,892 video moments, and 40,543 localized descriptions. In the dataset, each video is broken into six five-second moments

and represented by s 6 * 4096 feature matrix, where each column represents a 4096-d tsn [19] feature of one moment. For language features, they adopted 300 dimensional dense word embedding Glove [12] to represent each word. They use LSTM to extract feature for sentences.

4.2 Evaluation Metric

To thoroughly measure our model and the baselines, we adopt "R@n, IoU=m" proposed by [6] as the evaluation metric. Where IoU means the intersection temporal length divide the union temporal length of proposal and ground truth. To be more specific, given a query, it is the percentage of top-n results having iou larger than m. In the following, we use R(n,m) to denote "R@n, IoU=m". This metric itself is on the query level, so the overall performan

$$R(n,m) = \frac{1}{N_q}\sum_{i=1}^{N_q} r(n,m,q_i) \quad (7)$$

ce is the average among all the queries.

where $r(n,m,q_i)$ is the recall for a query q_i, N_q is the total number of queries, and R(n,m) is the averaged overall performance.

4.3 Baselines.

We compared our proposed MLN with the following several stateof-the-art baselines to justify the effectiveness of our proposal:

ACRN[9] This method jointly characterizes the attentive contextual visual feature and the cross-modal feature representation and is the first to introduce a temporal memory attention network to memorize the contextual information for each moment.

TALL This is a cross-modal temporal regression localizer that jointly captures the interaction between the query description and video moments, as well as outputs alignment scores and action boundary regression results for the moment candidates

MCN This method is designed for the moment-query retrieval task. It emphasizes the local and global moment features, aiming to strengthen the expressiveness ability.

VSA-RNN This method is the variant of the Deep VisualSemantic Alignment (DVSA) model [7]. It transforms the local visual feature and the texture feature encoded by the LSTM model into a common space, and then estimates the matching score of each moment candidate and the query.

4.4 System Variants

We experimented with variants of our system to test the effectiveness of our method. MLN-base: Our base model and follows naive training sample selection approach in section 3.2. MLN-aug: Our base model and implement samples augmentation in section 3.3. MLN-es: Our base model and implements effective samples selection in section 3.4. MLN: Our base model and implement both samples augmentation and effective samples selection.

4.5 Results and Comparison

In this part, we show and analyze the experiment results on TACoS and DiDeMo. For fair compare, we use visual features and sentence embedding provided from the datasets, so each methods share same features. We use original train, val, test splits. All the models are selected on validation set and report results on test set.

We test our system variants and baseline methods on TACoS and report the result for IoU of {0.1, 0.3, 0.5} and Recall@{1, 5}. The results are shown in Table 1. "Random" means that we randomly select n windows from the test sliding windows and evaluate Recall@n with IoU=m. VSA-RNN uses the end-to-end trained LSTM as the sentence encoder and all other methods use pretrained Skipthought as sentence embedding extractor. We can observe that:

MCN performs the poorest in TACoS, and is just a little better than random. This is because MCN use average feature of whole videos as context feature. It works well in notso-long videos(videos of DeDiMo are 25-30s), but for long videos(videos of TACoS are several minutes), there are too much noisy features.

VSA-RNN perform much better than MCN. It captures correct feature of videos and languages. Its gaps with TALL is due to :(1)it calculates similarity without embedding features

Table1. ComparisonofourMLNvariantsandothermethods on TACoS. Where in "R@n, I=m", "I" means IoU.

Method	R@1 I=0.5	R@1 I=0.3	R@1 I=0.1	R@5 I=0.5	R@5 I=0.3	R@5 I=0.1
Random	0.83	1.81	3.28	3.57	7.03	15.09
MCN	1.25	1.64	3.11	1.25	2.03	3.11
VSA-RNN	9.96	16.16	20.92	18.32	29.19	40.66
TALL	12.46	16.85	21.69	24.44	33.38	45.38
ACRN	14.62	19.52	24.22	24.88	34.97	47.42
MLN-base	16.2	20.9	25.9	27.9	37.4	47.7
MLN-aug	18.2	21.2	24.5	29.8	37.0	46.8
MLN-es	18.8	23.7	29.2	29.1	39.2	50.9
MLN	20.0	23.9	29.1	31.5	41.5	53.3

Table 2. Comparison of MLN and other methods on DEDiMo. Where in "R@n, I=m", "I" means IoU.

Method	R@1 I=0.5	R@1 I=0.7	R@1 I=0.9	R@5 I=0.5	R@5 I=0.7	R@5 I=0.9
MCN	23.33	15.37	15.32	41.03	20.37	19.77
VSA-RNN	24.94	14.52	14.44	68.39	26.10	23.95
TALL	26.45	15.36	15.31	68.78	28.43	26.15
ACRN	27.44	16.65	16.53	69.43	29.45	26.82
MLN	33.7	18.78	18.73	63.37	42.71	41.61

of different modal into a common space. (2)Predefined actions and objects are not precise enough to retrieval video clips.

ACRN outperforms MCN, VSA-RNN and TALL. By memorizing the context information and employing the attention mechanism on identifying the adaptive importance attention of each context moment, ACRN utilizes the contextual feature properly and it is beneficial in moment sentence localization.

Our base model outperforms all other previous methods. And our data augmentation and selection methods shows effectiveness.

We also evaluate our MLN model and the baseline methods on DiDeMo, and reported the results regarding IoU of {0.5, 0.7, 0.9} and Recall@{1, 5}. Note that since the positive moment-query pairs in this dataset are well aligned, we only used the alignment loss to train the ACRN ,TALL and MLN model for localizing the

corresponding moment. Results are shown in Table 2. We can see it that the results are consistent with those on TACoS. MLN shows a significant improvement over all models.

Figure(3) show an example result of different method. As we can see: (1)In the first picture, ARCN can not distinguish precise actions in matches while MLN can. (2)In the second picture, ARCN can not distinguish precise objects in matches while MLN can. It proves that MLN can capture video and language features more accurately. (3)In the third picture, MLN can not only retrieve simple query but also complex query. Figure(4) shows the recall rate with respect to top n. We can see it that recall increases as n increases and recall saturates when n is near 100. This shows the upper bound of MLN. And we believe the upper bound is due to proposal and refine procedure, and better proposal method will help. In general, we

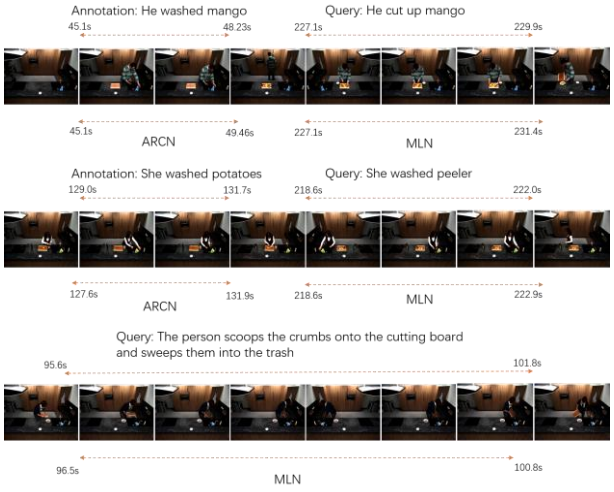


Figure 3. Results of MLN and ARCN in TACoS dataset. We find the annotation that matches the first proposal of both models most. Descriptions above images are the annotated sentences and their location. Descriptions below images are the clips proposed by models.

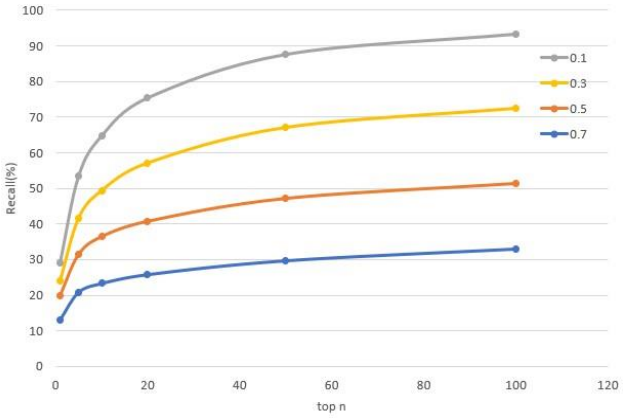


Figure 4. Recall curve on diferent IoU of MLN.

observed three types of common hard cases: (1) Clips that matches a long sentence. (2) Objects or actions of a clip only appear in test set and are unseen before. (3) Objects or actions are obscure. MLN can not distinguish plate and cutting board well.

5. CONCLUSIONS

We propose MLN model that can perform better on moment retrieval problem. Our novel data augmentation method can augment a lot of data and helps to fit the model. Our sample selection method can select the most effective samples for model training and can reduce overfit and performs better. In experiments, we observed three common hard cases: long sentence, unseen word and obscure objects. We think a better language model to capture sentence feature special for moment retrieval is needed. And although our sample method can use sample more sufficiently, a large dataset for moment retrieval can make a huge difference.

6. REFERENCES

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised Learning from Narrated Instruction Videos. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), 4575–4583.

[2] Piotr Bojanowski, Rémi Lagugie, Edouard Grave, Francis R. Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. 2015. Weakly-Supervised Alignment of Video with Text. 2015 IEEE International Conference on Computer Vision (ICCV) (2015), 4462–4470.

[3] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep VisualSemantic Embedding Model. In NIPS.

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ramakant Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. 2017 IEEE International Conference on Computer Vision (ICCV) (2017), 5277–5285.

[5] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing Moments in Video with Natural Language. 2017 IEEE International Conference on Computer Vision (ICCV) (2017), 5804–5813.

[6] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural Language Object Retrieval. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), 4555–4564.

[7] Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015), 3128–3137.

[8] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014. Visual Semantic Search: Retrieving Videos via Complex Textual Queries. 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014), 2657–2664.

[9] Meng Liu, Xiuli Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In SIGIR.

[10] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016), 11–20.

[11] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2016. Learning Joint Representations

of Videos and Sentences with Web Image Search. In ECCV Workshops.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In EMNLP.

[13] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding Action Descriptions in Videos. TACL 1 (2013), 25–36.

[14] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script Data for Attribute-Based Recognition of Composite Activities. In ECCV.

[15] Ozan Sener, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Unsupervised Semantic Parsing of Video Collections. 2015 IEEE International Conference on Computer Vision (ICCV) (2015), 4480–4488.

[16] Richard Socher and Li Fei-Fei. 2010. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2010), 966–973.

[17] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. TACL 2 (2014), 207–218.

[18] Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning LanguageVisual Embedding for Movie Understanding with Natural-Language. CoRR abs/1609.08124 (2016).

[19] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In ECCV.

[20] Ran Xu, Caiming Xiong, Wei Chen, and Jason J. Corso. 2015. Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework. In AAAI.

A New Multi-Camera Dataset with Surveillance, Mobile and Stereo Cameras for Tracking, Situation Analysis and Crime Scene Investigation Applications

Thomas Pollok
Fraunhofer IOSB
Fraunhoferstr. 1
76131 Karlsruhe, Germany
thomas.pollok@iosb.fraunhofer.de

ABSTRACT

Nowadays, cameras are ubiquitous. In the event of a crime or terrorist attack, plenty of video data can be collected from surveillance cameras or mobile phone cameras from witnesses. This paper presents a new dataset, "IOSB-4D," which includes data from multiple static and mobile cameras (from potential witnesses) that record a site. We provide several scenarios involving 10 actors that contain scenes of kidnapping, attack of a person by a group, loading a briefcase into a trunk, and more. In total, there are recordings from three perspectives from stereo cameras as well as two static surveillance cameras and up to six mobile cameras. For all cameras, we provide intrinsic parameters, and for the static cameras, we provide 6DOF pose parameters. Furthermore, high resolution images of all actors are available. This new dataset is valuable for object tracking and situation analysis, or research on innovative visualizations of combined camera and video footage, for example, to create a visualization in 3D/4D for a post analytic crime scene investigation.

CCS Concepts

• Computing methodologies~Camera calibration • Computing methodologies~3D imaging • Computing methodologies~Activity recognition and understanding • Computing methodologies~Scene anomaly detection

Keywords

Dataset; Multi-Camera Calibration; Crime Scene Investigation; Stereo; Mobile Camera; Surveillance Camera.

1. INTRODUCTION

Cameras are widely used for various reasons, from video surveillance to simply taking a photo or video of an interesting event, like a marathon, with a mobile phone. In case of a crime or terrorist attack, all of this data can be valuable evidence for a crime scene investigation. The goal of this paper is to provide researchers with a new high resolution multi-camera dataset that enables the development of tracking and situation analysis approaches on the basis of timestamp synchronized video data. Furthermore this dataset should enable researchers to work on post analytic visual exploration of crime scenes in 3D and 4D based on a large amount of video data from different perspectives. Our dataset includes image sequences from surveillance cameras, mobile phone footage, and stereo cameras from which depth maps can be computed. A three-dimensional representation of all fixed cameras can be achieved with this dataset, as we provide a complete set of intrinsic camera parameters together with full 6DOF multi-camera pose calibration.

In literature, there are various datasets for tracking and situation analysis tasks [1][2][3]. "Dana36," as described in [1], provides a dataset with multiple camera views that contains an annotated ground truth consisting of a region of interest together with an ID of all persons and cars. "3DPeS" (3D People Dataset for Surveillance and Forensics) [2] provides a fully calibrated multi-camera dataset with a modeled 3D scene. It contains annotations like bounding boxes with person IDs and trajectories.

However, none of the datasets in literature fully address calibrated multi-camera setups with stereo cameras that allow obtaining depth maps for static camera views that can be used for 4D scene representation. Furthermore, the data presented in literature only contains views from fixed cameras, like surveillance cameras, but does not provide views from moving mobile cameras, like footage recorded by potential crime witnesses.

This paper is structured into three sections. First is an overview of the dataset with short descriptions of all six recorded scenarios. Afterwards, the camera setup and the acquired images are described. Finally, a conclusion is presented.

2. Description of the Dataset

2.1 Overview

Our dataset, "IOSB-4D", was recorded in the courtyard of Fraunhofer IOSB, in Karlsruhe, Germany, in February 2018. In total, six different scenarios were recorded to illustrate the use of the dataset for things like tracking, re-identification and abnormal behavior detection. For recording, we used multiple distributed cameras that include three stereo cameras, two static surveillance cameras and multiple mobile cameras from various smartphones. This dataset provides only the raw video data with camera parameters, but no further ground truth annotation like object or person bounding boxes. An overview of the scene and the camera locations is presented in Figure 1. The red dotted bounding box highlights the area of interest. Two red triangles denote the positions and orientations of the static surveillance cameras of which one is mounted on the facade of the building and another one on the pole of a street lamp. The yellow triangles show the

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301542

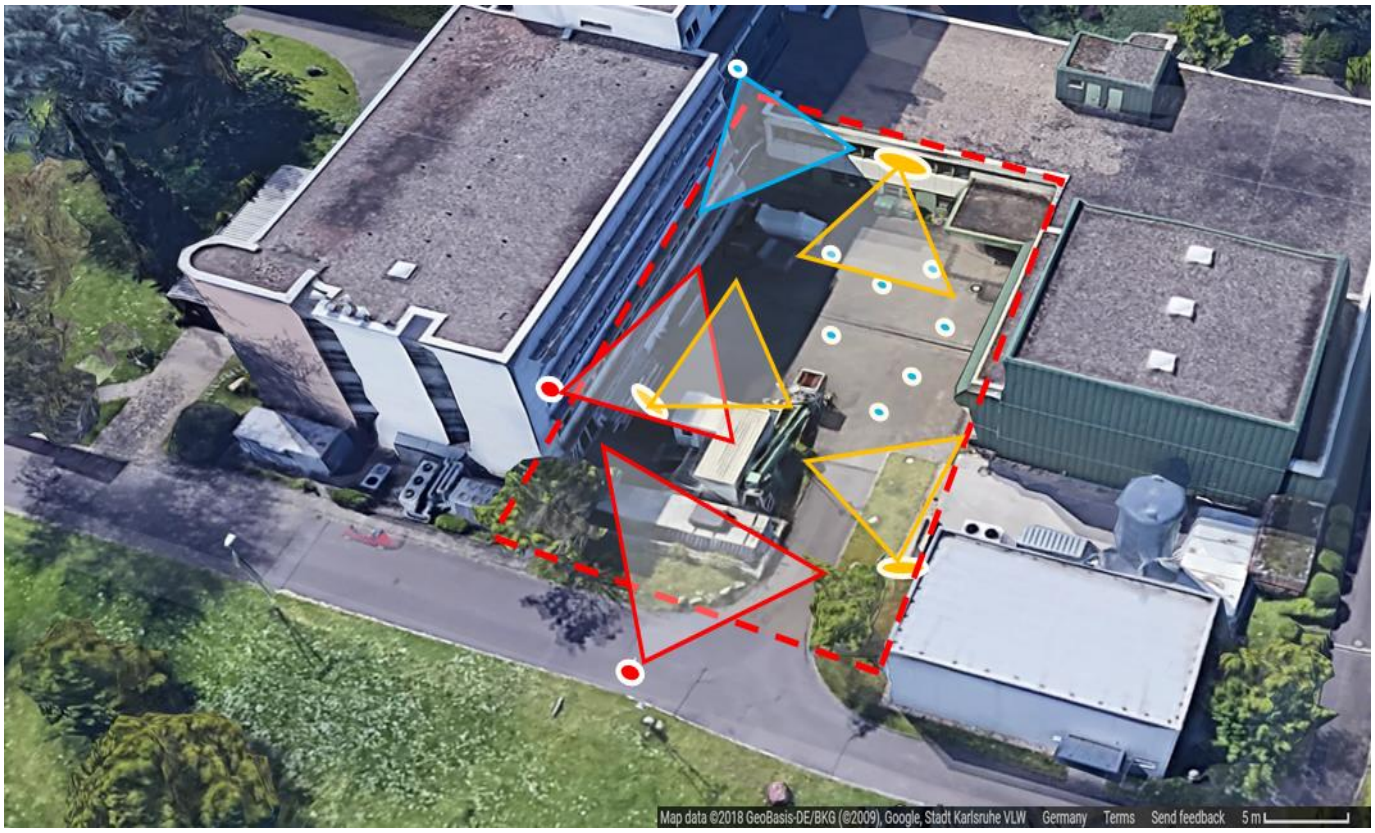


Figure 1. Overview of the scene covered by the proposed dataset. The positions and orientations of the static surveillance cameras are highlighted in red, the positions and orientations of the static stereo cameras are highlighted by yellow triangles and the blue dots show the approximate locations of the actors with their mobile cameras.

locations of the three stereo cameras, which are oriented towards the center of the scene. Blue dots depict the rough locations of the participants who were recording the scenarios with their mobile smartphone cameras. One participant was recording the scene from the inside of the building and is highlighted by the blue triangle.

2.2 Scenarios

Our dataset consists of six different scenarios in total. Each of them is stored as a separate dataset with a duration of approximately thirty to sixty seconds, focusing on scenes with actual action.

Scenario 1: "Item exchange"

Two people walk towards each other and meet in the center of the courtyard. They exchange an item and leave the scene.

Scenario 2: "Unattended briefcase"

Pedestrians walk in the courtyard while two people carry a briefcase. One of them puts the briefcase onto the ground, opens and closes it and leaves without the briefcase. Curious pedestrians walk towards the briefcase and film it with their cameras.

Scenario 3: "Kidnapping"

Two small groups of people are having a conversation in the courtyard. A car enters the scene and two unknown men get out of the car in order to kidnap a person from one of the groups. They force the hostage into the car and the car drives away. Nobody helped the person to escape, but gazers recorded videos with their mobile phones.

Scenario 4: "Group attack"

A pedestrian walks into the courtyard and immediately gets attacked by a group. Another group realizes the attack is underway and tries to help the victim so that he can run away. Some gazers are recording videos with their phones.

Scenario 5: "Briefcase gets loaded into a trunk"

A car drives into a courtyard and stops. Two suspects walk with a briefcase towards the car's trunk, open it and load the briefcase into the trunk. The car drives away, but gazing pedestrians have filmed the scene. Figure 3 shows different views of this scenario from all three stereo cameras, surveillance and mobile cameras at the same point in time.

Scenario 6: "High dynamics"

Pedestrians are walking at different velocities in the scene and occlude each other temporarily. The scene also contains also a car that drives slowly through the crowd.

3. CAMERA SETUP

3.1 Stereo Cameras

To allow for estimating dense depth maps for each captured point in time, we recorded the scene using three custom built stereo cameras with varying stereo baselines and focal lengths of 8.5mm, 12.5mm, and 16mm. All sensors have an image resolution of 1920x1200 pixels. Two stereo cameras use grayscale sensors while one stereo camera is equipped with a RGB sensor. For recording stereo image pairs, we used an external trigger system

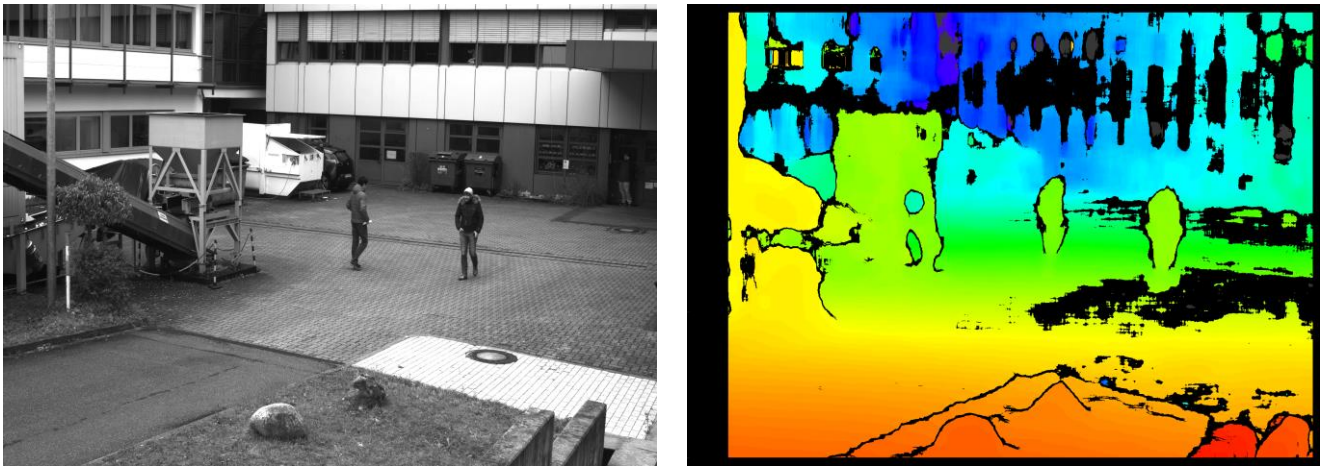


Figure 2. Depth map estimated from a stereo camera image pair using stereo block matching from the OpenCV library [6].

that allowed us to request images at the same point in time from the cameras. Regardless of the fact that all six cameras of the three stereo cameras were connected to the external trigger system, only the image pairs per stereo camera are synchronous. The image index of one stereo camera does not allow for direct access of the same time point from a different view since dropped frames may occur during the recording. However, the frames of all cameras are synchronized by a timestamp encoded in the file name. All three stereo cameras were calibrated separately in terms of intrinsic camera parameters as well as stereo calibration using a chessboard calibration pattern [4][5]. The estimated parameters are provided with our dataset, allowing for reconstructing depth maps from stereo image pairs. Furthermore, all cameras are calibrated to a common reference coordinate space which allows for visualizing all depth maps of a corresponding time point in a point cloud representation. For the reconstruction of the depth maps from the provided stereo data, we used an implementation of the stereo block matching algorithm provided with the OpenCV library [6]. An exemplary result of a depth map is shown in Figure 4.

3.2 Surveillance Cameras

Two surveillance cameras were mounted in typical surveillance camera locations for a courtyard scene. The first was mounted on the façade of the building, and the second was mounted on the pole of a street lamp. Both cameras were facing towards the courtyard and were recording images during all scenarios. In the dataset we provide the camera images with timestamps encoded in the filenames as well as the extrinsic calibration relative to the common coordinate space with the stereo cameras and intrinsic camera parameters. The image resolution of the cameras is 1920x1080 pixels (FullHD). Two exemplary views of the cameras are shown in Figure 3 (second row).

3.3 Mobile Cameras

The participants were instructed to record video footage with their standard mobile phone cameras during the scenarios. In each scenario, only a subset of all participants were recording video sequences while the remaining participants were the actual actors. The motivation was to collect mobile camera videos that observe the scene from the human perspective, since in real-world scenarios or actual incidents witnesses typically record evidence using their mobile phone. The videos are not pre-processed for this dataset, meaning that camera pose estimation per frame is not provided by this dataset. However, the intrinsic camera parameters are provided and all videos contain time stamp meta data. The provided mobile camera image sequences can also be used for scene reconstruction based on Structure-from-Motion (SfM) approaches. Figure 3 shows the result of a dense SfM-based scene reconstruction from one of the videos using COLMAP [7]. The dense point cloud can be used as a static reference point cloud of the scene, while the stereo depth data from the stereo cameras and all view frustums can be mapped into the static point cloud for visualization.

3.4 Actor Images

Additionally to the image and video sequences recorded from all cameras, high resolution frontal images of the participants are provided with this dataset to allow for tracking and re-identification applications of individuals. An overview of all individuals is shown in Figure 3 (bottom). In total, 10 actors were involved in this dataset. Note that not every actor is always visible in each scenario.

4. CONCLUSIONS

In this paper, we present a new dataset, "IOSB-4D," that consists of six surveillance scenarios. Our main contributions are the availability of time stamp synchronized high resolution (FullHD) images from multiple distributed cameras of which three are stereo cameras; a complete set of intrinsic and extrinsic (6DOF) camera parameters of all static cameras; and high resolution images of all involved actors. We think that this new dataset is valuable for the development of visual tracking and situation analysis algorithms. The availability of multiple stereo cameras and the camera parameters dataset allow for research on innovative combined interactive visualizations of all cameras and video footage, for example, in 3D/4D for a post analytic crime scene investigation. To promote academic research in the related areas, the dataset will be made publicly available to research purposes upon request.

5. ACKNOWLEDGMENTS

This study was partially supported by the FLORIDA project, co-funded by the German Federal Ministry of Education and Research (BMBF) under grant 13N14251.

Figure 3. Views from the different cameras at the same point in time from scenario 6. Top: stereo cameras. Second row: surveillance cameras. Third row: mobile phone footage. Bottom: Frontal images of all involved actors.

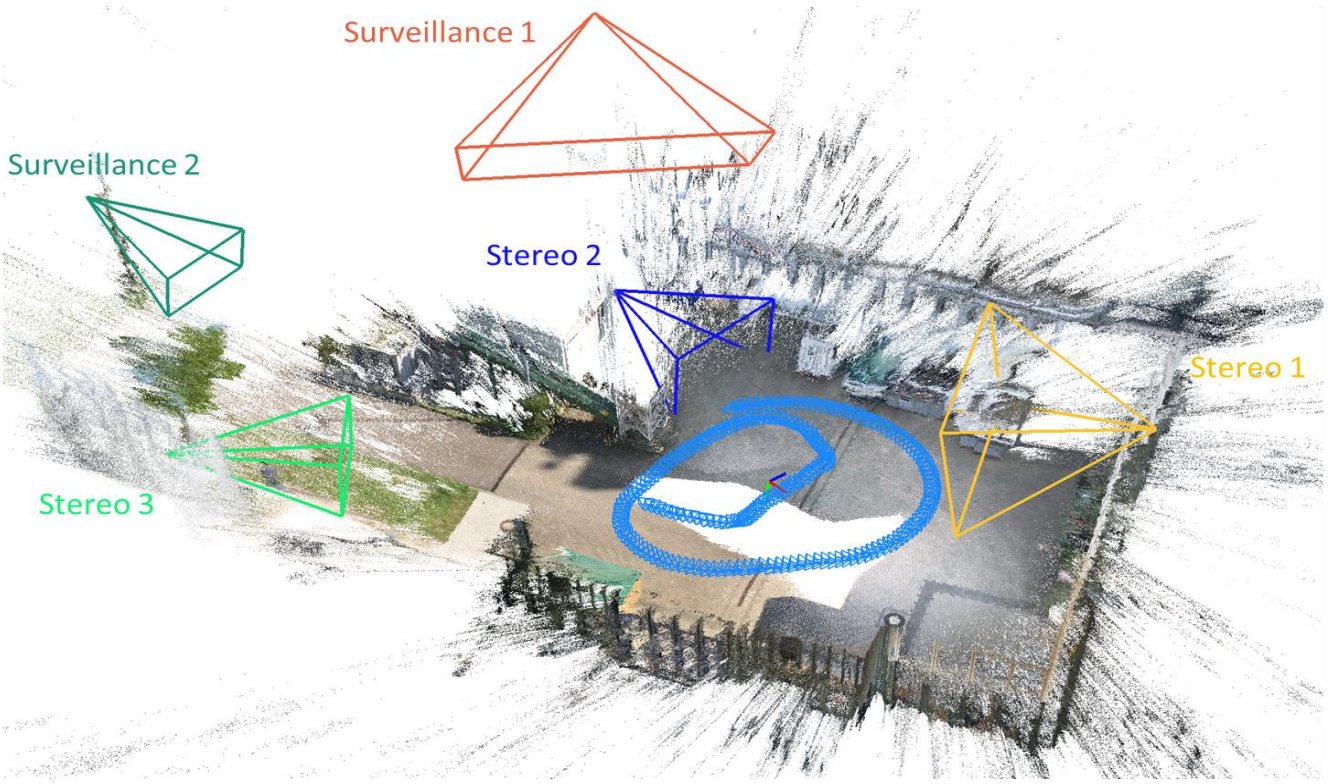


Figure 4. Dense static scene reconstruction of the scene based on a mobile camera scene video using COLMAP [7] with all stereo cameras and surveillance cameras registered to the scene. The light blue track of view frustums show the path of the sampled mobile camera frames used for the dense static scene reconstruction.

6. REFERENCES

[1] J. Per, V. S. Kenk, R. Mandeljc, M. Kristan and S. Kovacic, "Dana36: A Multi-camera Image Dataset for Object Identification in Surveillance Scenarios," *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, Beijing, 2012, pp. 64-69.

[2] D. Baltieri, R. Vezzani, R. Cucchiara. "3DPes: 3D People Dataset for Surveillance and Forensics". in *Proceedings of the 1st International ACM Workshop on Multimedia access to 3D Human Objects*, Scottsdale, Arizona, USA, pp. 59-64, Nov 28 - Dec 1, 2011.

[3] A. Nambiar, M. Taiana, D. Figueira, J. Nascimento, A. Bernardino, "A multi-camera video dataset for research on high-definition surveillance". Int. J. of Machine Intelligence and Sensory Signal Processing, 2014 Vol.1, No.3, pp.267 - 286

[4] A. De la Escalera and J. M. Armingol. "Automatic chessboard detection for intrinsic and extrinsic camera parameter calibration". Sensors, 10(3):2027, 2010.

[5] Z. Zhang. "A flexible new technique for camera calibration". IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11):1330–1334, Nov 2000.

[6] G. Bradski. "The OpenCV Library". Dr. Dobb's Journal of Software Tools, 2000.

[7] J. L. Schönberger, J.-M. Frahm. "Structure-from-Motion Revisited". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 4104-4113.

A Fast Block Partitioning Algorithm Based on SVM for HEVC Intra Coding

Jun Yin
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
yin_jun@dahuatech.com

Xu Yang
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
yang_xu3@dahuatech.com

Jucai Lin
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
lin_jucai@dahuatech.com

Yao Chen
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
chen_yao@dahuatech.com

Ruidong Fang
Zhejiang Dahua Technology co.,ltd
No.1199 Bin'an Road,
Binjiang District, Hangzhou, China
+86-571-28932955
fang_ruidong@dahuatech.com

ABSTRACT
Quad-tree-based coding unit (CU) block partitioning structure in High Efficiency Video Coding (HEVC) intra prediction causes a significant coding complexity increase. Hence, a fast block partitioning algorithm based on Support Vector Machine (SVM) is proposed in this paper. Firstly, some effective features are extracted from CUs in each depth as the input vector of SVM. Secondly, three offline trained SVM CU splitting models are loaded in each CU depth, which predict the class label of the current CU according to the extracted features. Moreover, the parameters of SVM models are resolved by grid search method. Finally, based on the predicted class label, the encoder will decide whether to split the current CU or not. Experimental results show that the proposed algorithm reduces the computational complexity of HM13.0 to 30.1% and 53.9% in encoding time with and without RDO(rate distortion optimization), while the loss in coding efficiency is negligible.

CCS Concepts
• Computing methodologies → Machine learning approaches

Keywords
High Efficiency Video Coding (HEVC), block partitioning, intra coding, Support Vector Machines (SVM), feature extraction

1. INTRODUCTION
Developed by the Joint Collaborative Team on Video Coding (JCT-VC) of ISO/IEC and ITU-T, High Efficiency Video Coding (HEVC) [1] aims to maintain the equal output video quality while reducing half of the bitrate requirement compared with H.264/AVC [2]. HEVC is also based on a hybrid coding framework [3] but introduce many new tools [4] to improve the coding efficiency, such as flexible coding structures include the transform unit (TU), prediction unit (PU), coding unit (CU), quad-tree-based CU block partitioning, and up to 35 intra prediction modes, etc. For selecting the best quad-tree partitioning of the coding tree unit (CTU), HEVC Test Model (HM) reference software uses rate distortion optimization (RDO) technique to find the best mode with the minimum rate distortion (RD) cost. However, the partitioning process costs a lot of time. The increased computational complexity limits the application of HEVC. Hence, reducing the computational complexity of the HEVC encoder without affecting its coding performance has become a hot topic in the field of video encoding.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301527

In this paper, we propose a fast block partitioning algorithm based on Support Vector Machines (SVM) for HEVC intra coding. The major contributions of this paper are as follows:

1. The key idea of the proposed algorithm is to utilize the offline trained SVM CU splitting model to early terminate the partitioning process. We regard CU partitioning as a binary classification problem, which can be classified into "split" and "non-split" classes.

2. We firstly extract some effective features from CUs in every depth as the input vector of SVM. Then, three offline trained SVM CU splitting models are loaded in each CU depth, which predict the class label of the current CU. Moreover, the parameters of the SVM model are resolved by grid search method. Finally, the encoder will decide whether to split the current CU based on the predicted label.

The remainder of this paper is organized as follows: Section 2 reviews the related work about fast block partitioning algorithms for HEVC intra coding. Section 3 presents a detailed description for the proposed fast block partitioning algorithm based on SVM for HEVC intra coding. Section 4 shows the experimental results. Finally, Section 5 concludes the paper.

2. BACKGROUND REVIEW
Recently, much attention has paid to fast block partitioning algorithms for HEVC intra coding. The existing methods can be categorized into two types: content-based methods and statistic-

based methods. In [5], Yao et al. proposed the most possible depth range and employed dominant edge detection (DEA) to predict the partitioning depth. In [6], Shen et al. employed the nearby tree blocks depth level to skip some depth levels. Cho et al. [7] proposed an early CU splitting method according to Bayes decision based on the full RD and the low complexity costs. In [8], Shang et al. presented a fast CU size decision algorithm for HEVC intra coding. It exploited the depth information of co-located CUs to make an early CU pruning decision or CU split decision. In [9], Wang et al. proposed a fast CU size decision algorithm which could improve the speed of intra coding in HEVC. With the statistical analysis for spatial correlation of images, they firstly classified the CUs into three cases according to different neighboring block partitions. Then, the CU size was determined early by referencing the neighboring CUs and combining with necessary comparison of luminance variances.

As can be seen from above, most of the researches focus on computation reduction. Therefore, the purpose of our proposed algorithm is to reduce computational complexity while maintaining coding efficiency as much as possible.

3. SVM BASED FAST BLOCK PARTITIONING ALGORITHM

HEVC supports four CU sizes for all CUs, ranging from 64 x 64 to 8 x 8, as shown in Figure 1. The CU can be recursively divided into four equal-sized sub-CUs until it reaches the minimum CU size. In order to choose the optimal CU partition size, HEVC encoder must compare the RD cost between the four sub-CUs and the CU in every depth level. Because of this, the computational complexity is excessively high. If we can predict the CU partitioning size before the RDO process, the computational complexity of HEVC intra coding can be reduced.

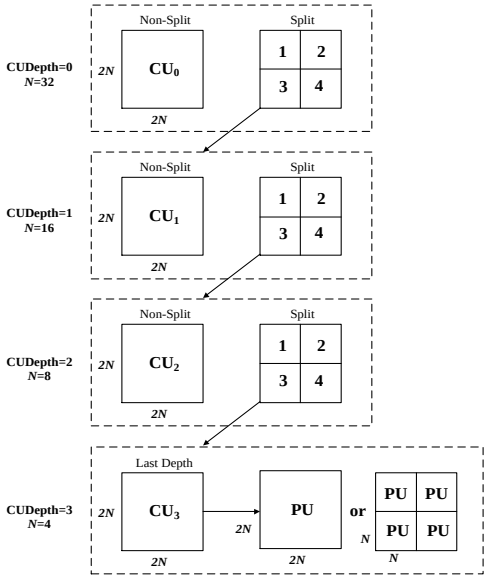


Figure 1. Illustration of recursive CU partitioning structure

3.1 Feature Selection

In general, the intra coding is intended to reduce the image spatial redundancy. It is clear that the homogeneous region has a higher probability to be coded as large blocks. Therefore, if we can calculate the complexity of the image in advance, some original CU splitting can be terminated early, which can speed up the intra RDO process and reduce the coding time.

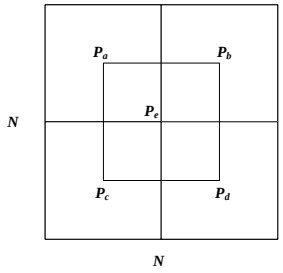


Figure 2. $N \times N$ size block partition into five $N/2 \times N/2$ size sub- blocks

In natural pictures, nearby CUs usually hold similar textures. Consequently, the optimal depth level of current CU may have a strong correlation with its neighboring CUs. Based on this concept, we exploit the maximum depth of the adjacent CUs as the feature to determine the current CU quad-tree partitioning. The maximum depth of the left CU, upper left CU, upper CU and upper right CU are named as x_{DL}, x_{DUL}, x_{DU} and x_{DUR}.

Variance of the image pixel is an important feature which can represent the image complexity. Thus, we employ the variance of the current CU as the feature to determine the CU quad-tree partitioning. The variance of the current CU is named as x_{VAR}, which is calculated as

$$x_{VAR} = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} [P(i,j) - \bar{P}(i,j)]^2 \quad (1)$$

where N represents the current CU size, and $P(i,j)$ and $\bar{P}(i,j)$ represent the luminance and the mean luminance values of the current CU, respectively.

Sobel, Roberts [10] and Canny [11] are the popular edge detection methods. Sobel and Roberts are simple to compute, but the detection results are not so good. Canny is complex to extract as feature, so we use DEA [12] to detect the edge information, which has a better detection performance. The DEA calculation steps are described in Figure 2. First, an $N \times N$ size block is equally divided into four $N/2 \times N/2$ size portions and an $N/2 \times N/2$ size central portion named a, b, c, d and e. The average pixel values of each portion named P_a, P_b, P_c, P_d and P_e are computed by

$$P_a = \sum_{i=0}^{\frac{N}{2}-1} \sum_{j=0}^{\frac{N}{2}-1} P(i,j) \Big/ \frac{N \times N}{4} \quad (2)$$

$$P_b = \sum_{i=\frac{N}{2}}^{N-1} \sum_{j=0}^{\frac{N}{2}-1} P(i,j) \Big/ \frac{N \times N}{4} \quad (3)$$

$$P_c = \sum_{i=0}^{\frac{N}{2}-1} \sum_{j=\frac{N}{2}}^{N-1} P(i,j) \Big/ \frac{N \times N}{4} \quad (4)$$

$$P_d = \sum_{i=\frac{N}{2}}^{N-1} \sum_{j=\frac{N}{2}}^{N-1} P(i,j) \Big/ \frac{N \times N}{4} \quad (5)$$

$$P_e = \sum_{i=\frac{N}{4}}^{\frac{3N}{4}-1} \sum_{j=\frac{N}{4}}^{\frac{3N}{4}-1} P(i,j) \Big/ \frac{N \times N}{4} \quad (6)$$

After getting the average pixel values P_a, P_b, P_c, P_d and P_e, the four DEAs (i.e., *DEA1*, *DEA2*, DEA3 and *DEA4*), corresponding to at 0°, 45°, 90°, and 135°, are given by

$$DEA1 = |P_b - P_a| + |P_d - P_c| \quad (7)$$

$$DEA2=\left|P_c-P_e\right|+\left|P_e-P_b\right| \quad (8)$$

$$DEA3=\left|P_c-P_a\right|+\left|P_d-P_b\right| \quad (9)$$

$$DEA4=\left|P_d-P_e\right|+\left|P_e-P_a\right| \quad (10)$$

Finally, we employ the four DEAs as the feature to determine the CU quad-tree partitioning, which are named as x_{DEA1}, x_{DEA2}, x_{DEA3} and x_{DEA4}.

In HEVC intra coding process, quantization parameter (QP) directly influences the partitions size. Thus, we add QP as a feature to determine the CU quad-tree partitioning. Here QP is named as x_{QP}.

3.2 SVM-based Two Classifier for CU

To reduce the computational complexity of CU partitioning, we regard CU partitioning as a binary classification problem. There are many ways to solve this problem. In this algorithm, we use the SVM model to resolve this issue. The main idea of SVM is to derive a separating hyper-plane which can maximize the margin between different classes. Given l training data points

$$\{x_i,y_i\}_{i=1}^{l},x_i\in R^N,y_i\in\{+1,-1\},i=1,2,\mathrm{L}\ ,l \quad (11)$$

where $\{x_i, y_i\}$ is the i-th training sample. x_i represents the input feature vector of each CU and y_i represents the class label indicating CU splitting or not splitting. In this paper x_i = {x_{DL}, x_{DUL}, x_{DU}, x_{DUR}, x_{VAR}, x_{DEA1}, x_{DEA2}, x_{DEA3}, x_{DEA4}, x_{QP}} and y_i ={+1,-1}. The SVM optimization problem is

$$\begin{aligned} &\min\frac{1}{2}w^Tw+C\sum_{i=1}^{l}\xi_i \\ &s.t.\quad y_i(w^T\phi(x_i)+b)\geq 1-\xi_i,\xi_i\geq 0 \end{aligned} \quad (12)$$

where w represents the weight vector. ξ_i is the slack variable. C is the penalty parameter of error, and the error rate is higher when C is smaller. $\phi(\cdot)$ is the kernel function.

By introducing the Lagrange method, (12) could be obtained by

$$\begin{aligned} &\min\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{l}\alpha_i \\ &s.t.\quad \sum_{i=1}^{l}\alpha_i y_i=0,C\geq\alpha_i\geq 0,i=1,2,\mathrm{L}\ ,l \end{aligned} \quad (13)$$

where α_i,α_j are Lagrange multipliers. $K(x_i, x_j) = \phi^T(x_i)\ \phi(x_j)$ is the kernel function.

3.3 SVM Model Parameter Optimization

The introduction of the kernel functions has made SVM model been widely used. The commonly used kernel functions are as follows: polynomial, sigmoid and radial basis function (RBF). Among the list, RBF has the least parameters which is expressed as:

$$K(x_i,x_j)=exp(-\gamma\left\|x_i-x_j\right\|)^2 \quad (14)$$

After selecting the kernel function, the next work is parameter optimization. The result of this step could greatly affect the SVM decision function. Among many parameter optimization methods, the grid search method [13] is the most popular one.

Grid search method is respectively taking M values in C and take N values in γ, for the $M \times N$ combination of (C, γ), then to estimate the learning accuracy. The computational complexity of parameter optimization in our algorithm is negligible because we trained offline SVM models. The value of M and N are both set to 21, so the range of C and γ are $[2^{-10},2^{10}]$. We employ the combination (C, γ) of 2^1 steps, then get the highest learn accuracy.

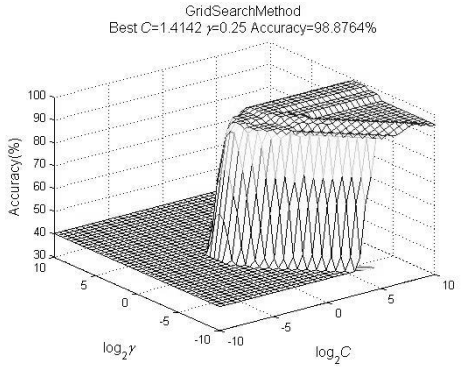


Figure 3. The accuracy rates of different parameters

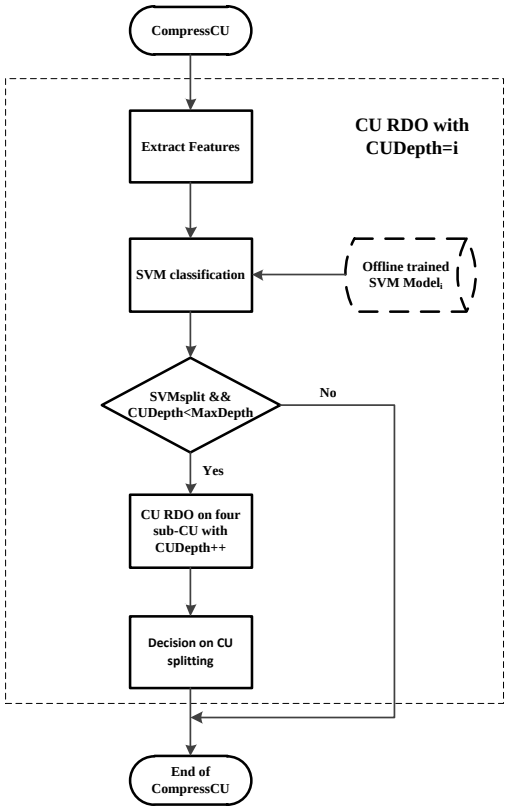


Figure 4. Block diagram of the proposed algorithm

As can be seen from Figure 3, we get the highest accuracy 98.8764% when C = 1.4142, γ = 0.25 by employing grid search method in our 0 level CU splitting model.

3.4 Process of the Proposed Method

The block diagram of the proposed algorithm based on SVM is shown in Figure 4. Firstly, feature vectors x_i = {x_{DL}, x_{DUL}, x_{DU}, x_{DUR}, x_{VAR}, x_{DEA1}, x_{DEA2}, x_{DEA3}, x_{DEA4}, x_{QP}} are extracted from CUs in every depth. Secondly, three offline trained SVM CU splitting models are loaded in each CU depth, which predict the class label y_i ={+1,-1} of the current CU according to the extracted features. Moreover, the parameters of the SVM model are resolved by Grid search method. Finally, based on the predicted class label, the encoder will decide whether to split the current CU or not.

4. EXPERIMENTAL RESULTS

We implement the proposed algorithm in the HM13.0 [14]. The first frame of four sequences (Four people, Kristen And Sara, People on street, Traffic, including smooth and complex textures) are used in SVM training for QP =22, 27, 32, 37. In the experiment, the parameters are set as All Intra (AI) configuration, and main profile including five classes of test sequences, QP values are 22, 27, 32 and 37. The system platform was the Inter Core i5-6500 CPU 3.20GHz. HEVC test sequences information is shown in Table 1. Because we extract only one frame features of the training sequences, so in the test stage, we also test the four training sequences for all frames. From above experiments, we can verify the generalization ability of our SVM model.

Table 1. Description of the HEVC test sequences

Sequences	Class	Num. of frames	Frame rate	Resolution (pixels)
PeopleOnStreet	A	150	30	2560×1600
Traffic	A	150	30	2560×1600
BQTerrace	B	600	60	1920×1080
BasketballDrive	B	500	50	1920×1080
Kimono	B	240	24	1920×1080
ParkScene	B	240	24	1920×1080
Cactus	B	500	50	1920×1080
BasketballDrill	C	500	50	832×480
RaceHorsesC	C	300	30	832×480
BasketballPass	D	500	50	416×240
BQSquare	D	600	60	416×240
RaceHorses	D	300	30	416×240
FourPeople	E	600	60	1280×720
KristenAndSara	E	600	60	1280×720
Johnny	E	600	60	1280×720

The coding efficiency is measured by the BD-Rate (%) and the BD-PSNR (dB) [15]. Moreover, the percentage difference in encoding time (ΔT) is also used to compare our algorithm with HM13.0. The criteria is calculated as

$$\Delta T = \frac{Time_{proposed} - Time_{HM13.0}}{Time_{HM13.0}} \times 100\% \quad (15)$$

4.1 Comparison with HM13.0

Table 2 shows the results of the proposed algorithm compared with HM 13.0. A false prediction of CU partitioning label usually leads to RD degradation, as the optimal RD performance is not achieved. To ameliorate this issue, when the CU is not split, we use the prediction label of SVM as the final partitioning result of current CU. While the CU is splitting, its four sub-CUs are all checked with the RD cost, which is the meaning of "with RDO" in table 2. Contrarily, "without RDO" means we just use SVM prediction results as the CU partitioning label, whether the prediction is right or not. The experimental results of "with RDO" show that the proposed algorithm reduces the computational complexity of HM13.0 to 30.1% in the encoding time, on average, while the encoding bit rate BD-Rate increases 0.85% and the BD-PSNR decreases 0.04dB. When "without RDO", the proposed algorithm reduces the computational complexity of HM13.0 to 53.9% in the encoding time, on average, while the encoding bit rate BD-Rate increases 1.27% and the BD-PSNR decreases 0.06dB.

As can be seen from Table 2, the proposed algorithm can reduce the coding time, whereas the coding efficiency loss is negligible.

Table 2. Results of the proposed algorithm compared with HM 13.0 standard algorithm

Sequences	With RDO			Without RDO		
	BD-Rate (%)	BD-PSNR (dB)	△T (%)	BD-Rate (%)	BD-PSNR (dB)	△T (%)
PeopleOn-Street	0.44	-0.02	-19.9	1.46	-0.08	-55.3
Traffic	0.92	-0.05	-24.1	1.06	-0.06	-53.7
BQTerrace	0.83	-0.05	-25.4	1.64	-0.10	-46.6
Basketball-Drive	0.89	-0.02	-39.6	2.01	-0.05	-53.1
Kimono	0.75	-0.02	-38.1	0.55	-0.02	-60.2
ParkScene	1.07	-0.05	-31.6	0.84	-0.04	-49.2
Cactus	0.91	-0.03	-29.0	0.72	-0.03	-50.0
Basketball-Drill	0.90	-0.04	-31.0	1.32	-0.06	-51.7
RaceHorsesC	0.91	-0.05	-27.0	1.05	-0.06	-52.2
Basketball-Pass	0.87	-0.05	-31.0	0.82	-0.05	-52.1
BQSquare	0.43	-0.03	-14.5	0.32	-0.03	-53.7
RaceHorses	0.92	-0.06	-20.2	0.85	-0.06	-48.5
FourPeople	1.23	-0.07	-29.8	1.93	-0.11	-56.8
KristenAnd Sara	0.79	-0.04	-43.7	2.20	-0.11	-63.2
Johnny	0.96	-0.04	-46.7	2.22	-0.09	-63.1
Average	**0.85**	**-0.04**	**-30.1**	**1.27**	**-0.06**	**-53.9**

In particular, for the "Kristen and Sara", and the "Johnny" sequences, the proposed algorithm saves much time because of these two sequences without complex texture. In this case, our algorithm can early terminate the process of partitioning, greatly saving the coding time.

4.2 Comparison with Other State-of-the-Art Work

Table 3 shows the performances of the proposed algorithm compared with [5] and [6]. Here we only use the "with RDO" results as reference to compare. The experimental results show that the coding performance and coding complexity of the proposed algorithm are better than that of the [5] and [6]. The proposed algorithm can save 6.4% coding time on average compared to [5] with 0.24% BD-Rate decrease. In particular, for the "BasketballDrive" sequence, the proposed algorithm can save 21.9% in coding time, while the BD-Rate decreases 0.18%. Additional, the proposed algorithm can save 5.6% coding time on average compared to [6], with a maximum of 15.9% in the "BasketballPass" sequence, while the BD-Rate decreases 1.01% on average.

We also select one frame in "BasketballPass" to show CU partitioning results of the proposed algorithm compared with HM

13.0. Figure 5 shows the CU partitioning results of HM13.0 in AI configuration. Figure 6 shows the CU partitioning results of the proposed algorithm in AI configuration. It can be seen from the figure that the CU partitioning result of the proposed algorithm is almost the same with HM13.0 standard algorithm.

Table 3. Results of the proposed algorithm compared with [5] and [6]

Sequences	[5]		[6]	
	BD-Rate (%)	ΔT (%)	BD-Rate (%)	ΔT (%)
PeopleOnStreet	-1.57	2.6	-1.93	1.7
Traffic	-0.03	-6.2	-1.27	-1.8
BQTerrace	-0.19	-3.3	-1.57	0.1
BasketballDrive	-0.18	-21.9	-2.15	-7.8
Kimono	-0.57	-17.2	-0.28	-2.1
ParkScene	0.00	-14.3	-1.14	-5.5
Cactus	-0.08	-12.1	-1.22	-5.5
BasketballDrill	-0.28	-12.1	-0.63	-13.1
RaceHorsesC	-0.04	-6.9	-0.53	-10.1
BasketballPass	-0.35	-0.5	-0.61	-15.9
BQSquare	-0.12	16.7	-0.60	0.4
RaceHorses	0.57	-1.9	-0.13	-7.3
Average	**-0.24**	**-6.4**	**-1.01**	**-5.6**

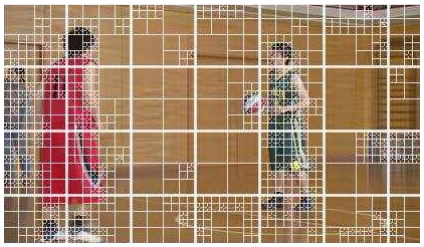


Figure 5. CU partitioning of HM13.0 in AI configuration

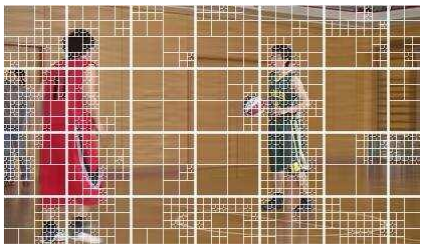


Figure 6.CU partitioning result of the proposed algorithm in AI configuration

5. CONCLUSION

A fast CU partitioning algorithm is proposed in this paper to reduce the computational complexity for HEVC intra coding. One of the important things is using the SVM splitting model to early terminate partitioning process. We firstly extracted some effective features from CUs in every depth, which is used to be the input vector of SVM. Then, three offline trained SVM CU splitting models are loaded in each CU depth, which predict the class label of the current CU according to the extracted input features. Moreover, the parameters of the SVM model are resolved by Grid search method. Finally, based on the predicted class label, the encoder will decide whether to split the current CU. Compared with HM13.0, the proposed algorithm can save 30.1% and 53.9% encoding time, with only a 0.85% and 1.27% BD-Rate increase and only a 0.04dB and 0.06dB BD-PSNR decrease, on average, for "with RDO" and "without RDO" case. Since our current work mainly focuses on the intra CU size selection, the inter CU size selection is worthy of future study.

6. REFERENCES

[1] Sullivan G J, Ohm J, Han W J, Wiegand T. Overview of the High Efficiency Video Coding (HEVC) Standard [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012: 1649-1668.

[2] Wiegand T (2003) Draft ITU-T recommendation and final draft international standard of joint video specification. ITU-T rec. H.264|ISO/IEC 14996-10 AVC.

[3] Wiegand T, Ohm J, Sullivan G J, et al. Special section on the joint call for proposals on high efficiency video coding standardization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(12):1661-1666.

[4] Bross B，Han W J，Sullivan G J，et al. High Efficiency Video Coding (HEVC) text spec, draft 10 (for FDIS&Consent) [S]. JCTVC-L1003，Geneva，Switzerland，2013..

[5] Yao Y, Jia T, Li X, et al. A fast DEA-based intra-coding algorithm for HEVC [J]. Multimedia Tools and Applications, 2017: 1-1.

[6] Shen L, Zhang Z, An P. Fast CU size decision and mode decision algorithm for HEVC intra coding. [J]. IEEE Transactions on Consumer Electronics, 2013, 59(1): 207-213.

[7] Cho S, Kim M. Fast CU Splitting and Pruning for Suboptimal CU Partitioning in HEVC Intra Coding [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2013, 23(9):1555-1564.

[8] Shang X, Wang G, Fan T, et al. Fast CU size decision and PU mode decision algorithm in HEVC intra coding [C]. IEEE International Conference on Image Processing. IEEE, 2015:207-213.

[9] Wang J, Dong L, Zhang J. A fast intra CU size decision algorithm based on spatial correlation in HEVC [J]. Journal of Computational Information Systems, 2015, 11(7):2605-2614.

[10] Suwanmanee S, Chatpun S, Cabrales P. Comparison of video image edge detection operators on red blood cells in microvasculature. Biomed Eng Int Conf (BMEiCON), Amphur Muang, 2013, 1–4.

[11] Canny J (1986) A computational approach to edge detection. IEEE Trans on Pattern Anal and Mach Intell PAMI 8(6):679–698.

[12] Tsai A-C, Wang J-F, Lin W-G, Yang J-F (2007) A simple and robust direction detection algorithm for fast H.264 intra prediction. IEEE Int Conf in Multimedia Expo, 1587–1590.

[13] L. Li, X. L. Zhang. "Optimization of SVM with RBF kernel".Computer engineering and Applications, 2006(29): 190-194.

[14] Kim I K, McCann K D, Sugimoto K, Bross B, Han W J and Sullivan G J. High Efficiency Video Coding (HEVC) Test Model 13 (HM13) Encoder Description. Document JCTVC-O1002, JCT-VC of ISO/IEC and ITU-T, Geneva, Switzerland, November 2013.

[15] G. Bjøntegaard. 2001. Calculation of average PSNR differences between RD-curves. Technical Report VCEG-M33, ITU-T SG16/Q6, Austin, TX, USA.

Chapter 6: Image Processing Technology and Application

High Signal to Noise Ratio Weld Pool Imaging Device Research in CMT+P

Pan Yang, Zhuang Zhao, Jing Han, Yi Zhang
School of Electronic and Optical Engineering, Nanjing University of Science and Technology
Nanjing, China
18851195990@163.com

Lianfa Bai
School of Electronic and Optical Engineering, Nanjing University of Science and Technology
Nanjing, China
blf@njust.edu.cn

ABSTRACT
A clear, arc-free weld pool image is the foundation for molten pool visual processing. In this paper, light intensity and welding current during CMT+P welding are captured synchronously through the photodiode and FPGA device. Through analyzing the light intensity and welding current, the CMT phase is detected by the current signal. And verified the accuracy by a high-speed digital camera. A suitable filter is realized by the FPGA to judge the CMT phase adaptively and trigger the camera to capture the weld pool images. A large mount online experimental results have demonstrated that this scheme can capture clear and high signal to noise weld pool images with almost no arc interference, which lays a solid foundation for weld pool vision based on CMT+P welding.

CCS Concepts
• **Hardware →Integrated circuits→ Reconfigurable logic and FPGAs→ Reconfigurable logic applications**

Keywords
weld pool visual processing，FPGA module，photodiode，high-speed digital camera，short circuit transfer

1. INTRODUCTION
In recent years, with the rapid development of modern process manufacturing, welding has become more and more important. Weld pool vision sensing is one of the key technologies for automation and intelligent control of welding processes. Most scholars have achieved certain results in sensing methods, image processing algorithms, and weld pool feature analysis[1-5]. However, it mainly focus on the image processing after the weld pool image is acquired, but little research on weld pool image capturing.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301536

MIG/MAG welding is the most widely used welding process. However, large heat input and inevitable splashing have limited its application in some fields, especially the thin plate less than 1mm is the restricted area for its application. As a new welding technology without slag splashing, CMT (Cold Metal Transfer) [6] has opened up a new field since small heat input, small deformation, no spatter and good bridging ability. It provides the perfect solution for the welding of thin plates. CMT technology provides a platform with the lower energy. Based on this, Fronius combines the CMT process with the pulse process to achieve a welding process named CMT+P(Cold Metal Transfer Mixed With Pulses). During CMT+P the CMT process and the pulse process are alternately mixed. i.e. after several pulse processes,the welding process convert into one or several CMT processes. There are two type of weld pool images acquisition: passive direct visual sensing[7-9] and active direct visual sensing[10]. Passive use the reflected arc light and the self-radiation of the weld pool to collect the pool image； however, active direct vision sensing use external light source illumination such as high-frequency laser, and use the composite filter system to obtain weld pool image. Since the active type requires an auxiliary light source, and its cost is much higher than the passive type, so passive vision sensing is more suitable for the actual production process. However, arc will strongly interfere the weld pool images in passive sensing.According to electrical parameters, an appropriate current trigger method can reduce arc interference and improve weld pool image quality. However, as far as we know, there are no systematic researches based on current trigger. Many papers simply propose to capture the image at the base current, without giving any specific selection criteria.

This paper focuses on the relationship between the welding current and light intensity in CMT+P welding process. Through analyzing the light intensity at different moments of CMT+P, FPGA is used to adaptively determine the CMT phase in the CMT process through the welding current signal. And verify its accuracy with a high speed camera. With trigger the camera at the CMT phase during CMT+P, weld pool images with almost no arc interference are captured, which provides a solid foundation to analysis weld pool image and welding quality[11-13].

2. EXPERIMENTAL DEVICE
The physical diagram of system device used in the experiment is shown in Figure 1. The FPGA model is Xilinx XC6SLX9. AD acquisition module is AD7607, its highest sampling frequency is 50KHZ. The photodiode is LSRSD-U1, and its response band is 200-1100nm. Through IV conversion amplifier, the change of light signal during welding will convert into voltage signal and measured by FPGA. The image acquisition device includes a BASLER acA 1920-155 um monochrome CCD camera, a

dimming film, and a V611 high-speed digital camera.When the resolution of V611 is 512 × 512, the maximum frame rate is 21000fps and the minimum exposure time is 1μs, which fully meets the experimental requirements. In the online welding experiment, the CMT+P welding process was used, the welding current was 140A, the welding speed was 24cm/min, and the welding power source used was FRONIUS CMT Advanced 4000Rnc.

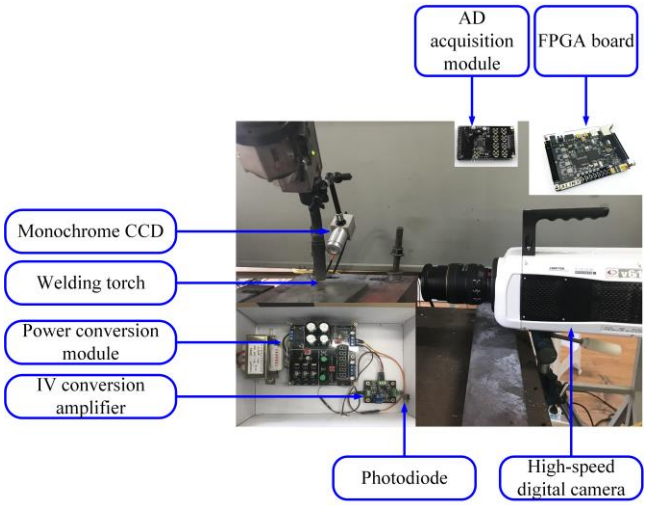


Figure 1. Physical diagram of system device.

3. SYSTEM DESIGN

The entire experimental system flow chart shown in Figure 2

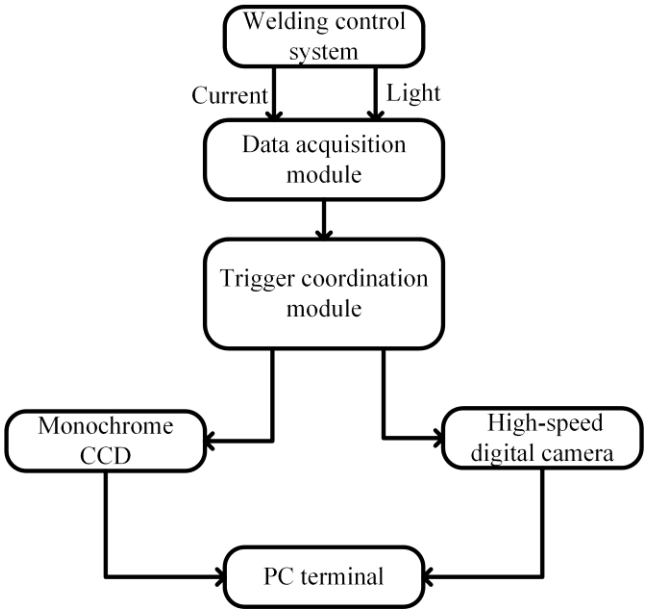


Figure 2. System flow chart.

The function module contains the signal acquisition module and the triggering cooperation module, verilog language is used to complete the programming design and simulation of the module function.

3.1 Signal acquisition module

The main function of the signal acquisition module[14] is to complete the synchronous acquisition of light intensity and current signal during online experiments, which mainly includes AD conversion module and SDRAM data storage module.

The AD7607 digital conversion module can simultaneously complete the acquisition of eight-way signals, and its sampling frequency can reach 50Khz, which fully meets the experimental requirements. In the experiment, ad_ch1 and ad_ch2 are used to separately collect the light intensity signal converted by the IV conversion amplifier and the current signal directly extracted by the welding control system. The capacity of the SDRAM data storage module is 256Mbit (16M*16bit), which can store the experimental data in the whole welding process. After the data acquisition is completed, the stored data in the SDRAM is exported through the serial port, thereby obtaining the light intensity and current data during the welding process.

Before performing the CMT+P welding experiment on the line, the signal acquisition module should be simulated to verify the stability of the acquisition module.Signal generator is used to provide a 100Hz sinusoidal signal with a 4V high level and 0V low level. FPGA is used to measure the sinusoidal signal.After downloading the running program, the data in the SDRAM is exported by the serial port, and the data signal diagram is shown in Fig. 3.

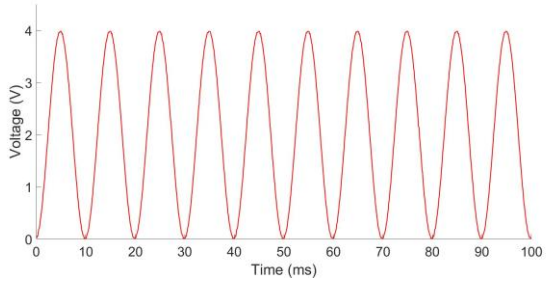


Figure 3. Simulation waveform diagram of signal acquisition module.

The period of measured signal is 10 ms and with 4V high level and 0V low level, respectively, which are the same as those provided by the signal generator. The signal acquisition module meets the design requirements and can be directly applied to online experiments.

3.2 Triggering collaboration module

According to the change of light intensity and current signal collected by the signal acquisition module, a suitable trigger program can be designed directly through the current signal. The trigger signal is given at the extinction time of the CMT, and the triggering cooperation module mainly includes the FILTER module, the TRIGGER module and the SDRAM data storage module.

The FILTER module is a simple digital low-pass filter[15] whose primary function is to filter out high current data while preserving low current data to ensure that the trigger signal is given during the current base phase. The collected current signal is saved in registers, and the number of registers is N (2<N<15), and the current values are respectively $I_1, I_2, \cdots\cdots, I_N$.When it satisfies the following conditions:

$$I_1 > I_2 > \cdots > I_N$$
$$I_1 - I_N >= I_C$$
$$I_N <= I_T \quad (1)$$

Where I_C is a different fixed value set for different welding current magnitudes, and the calculation formula of threshold I_T is:

$$I_T = \left(I_{tmin} + I_{Pmin}\right)/2 \quad (2)$$

The main function of the signal acquisition module is to complete the synchronous acquisition of light intensity and current signal during online experiments, which mainly includes AD conversion module and SDRAM data storage module..

I_{tmin} and I_{Pmin} are the CMT phase current minimum value and the pulse phase current minimum value, respectively. When formula (1) is established, the effect of filter is reached.

The function of the TRIGGER module is to detect the rising edge of the filtered signal and then give the trigger signal to control the CCD camera to capture the weld pool image.

The SDRAM data storage module stores the current value at the trigger time, that is, a current data I_{N+1} after I_N. So that the current value can be in one-to-one correspondence with the image captured at the triggering time, and verify the accuracy of the triggering time.

Before performing the CMT+P welding experiment on the line, it is necessary to simulate the triggering synergy module.In order to simulate the FILTER module and the TRIGGER module respectively. First, for the FILTER module, the signal generator is used to provide the same sinusoidal signal as the signal acquisition module simulates, and then derive the resulting data to obtain the signal diagram shown in Figure 4. For the TRIGGER module, a 100Hz square wave with a 4V high level and 0V low level is provided. After the simulation test, the signal waveform is shown in Figure 5.

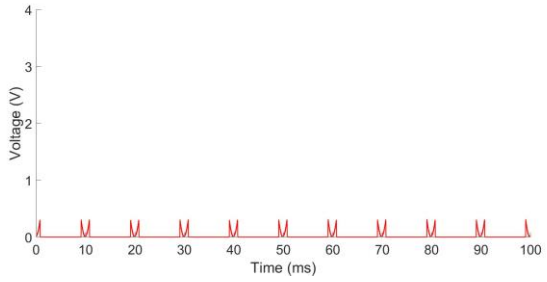


Figure 4. Simulation waveform diagram of FILTER module.

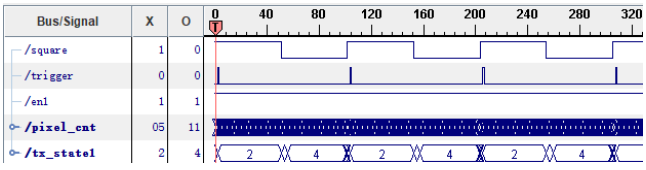


Figure 5. Simulation waveform diagram of TRIGGER module.

After analyzing the data in Figure 4, it can be seen that the FILTER module successfully filtered out all the high level data, retaining the low level data, indicating that the FILTER module function verification is successful.In figure 5, "square" is the square wave signal of the input analog, "trigger" is the trigger signal, the reference frequency is 1000Hz, after detecting the high level of square, a trigger signal appears, which indicates that the TRIGGER module successfully detected the rising edge and verify the module functional.

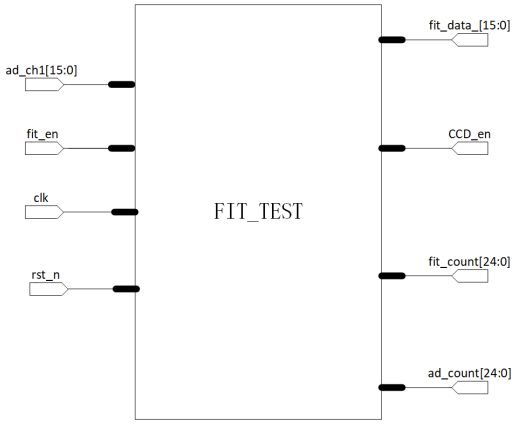


Figure 6. TRIGGER module RTL level diagram.

As shown in Figure 6, the RTL level diagram of the TRIGGER module is integrated. After the electrical parameter ad_ch1[15:0] passes through the FIT_TEST filter module, the pulse part of the electrical parameters are completely filtered out, leaving only the weaker electrical parameters in the CMT phase. Then, when the rising edge is detected, the trigger signal CCD_en is given at the rising edge time. The output data is set to fit_data[15:0]. This data is the electrical parameter data of the triggering time. It will enter the DDR3 and go online and output through the serial port to form a one-to-one correspondence with the collected molten pool image. ad_count[24:0] is the counter of the electrical parameter ad_ch1 entering the module, and fit_count[24:0] is the counter of the electrical parameter fit_data of the output module. Their functions are all for the subsequent verification of the molten pool image and the electrical parameters.

4. EXPERIMENTAL DESIGN AND ANALYSIS

After simulating test for the signal acquisition module and the trigger coordination module, online CMT+P welding experiment can be performed. The whole experiment is divided into three stages. Firstly, the data of the light intensity and current signal during the welding process are collected, and the light intensity signal is analysed to obtain the appropriate triggering time. Then the appropriate triggering procedure is used and verified its accuracy through the V611 high-speed digital camera. Finally, weld pool images with high signal to noise ratio are directly acquired by the TRIGGER module and the FILTER module.

4.1 Signal data collection and analysis

The normalized synchronously light intensity and welding current during the welding process, normalized are shown in Figure 7. It is obvious that the welding process has entered a CMT phase after six pulse stages, and the light intensity and current reach the lowest value in the CMT stage. During the pulse phase, the relations between light intensity and welding current are change synchronously, however, during the CMT phase are not synchronous, in the CMT phase, when the current rises slightly, the light intensity reaches its minimum. That is to say, it is arc-extinguishing time during the welding and we can capture weld pool images with high signal to noise ratio at this moment.

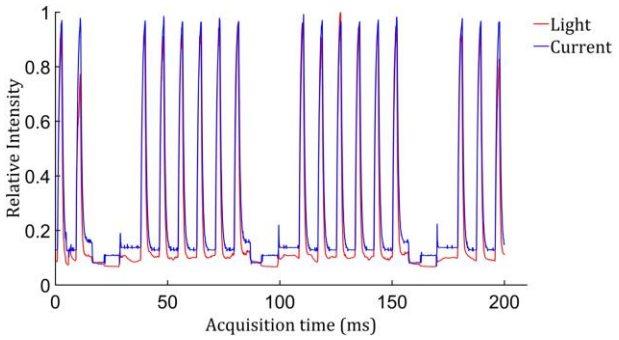


Figure 7. Current and intensity of light.

4.2 Shooting and verification of high frame rate weld pool images

After analyzing the relationship between light intensity and welding current, the accuracy is further verified through the V611 high-speed digital camera.When performing the online welding experiment, FPGA is used to give a 10KHz clock signal, and the TRIGGER module is used to collect the current data at the rising edge of the signal, and the triggers V611 high-speed camera to capture the weld pool image. The frame rate is 10000fps, which is lower than the maximum 21000fps. In this way, the synchronous acquisition of the current data and the weld pool image is realized, and the corresponding weld pool image can be accurately found at each stage of the welding.

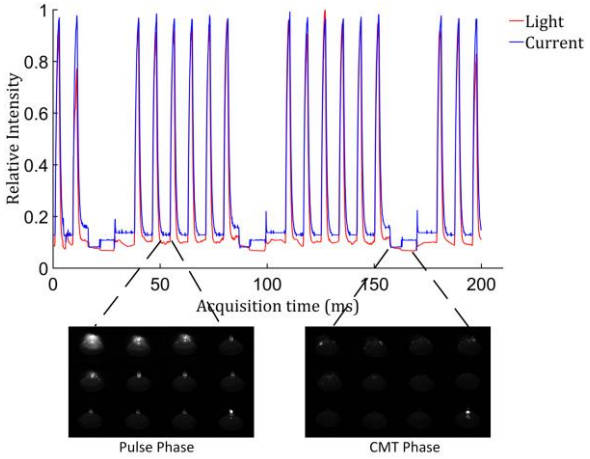


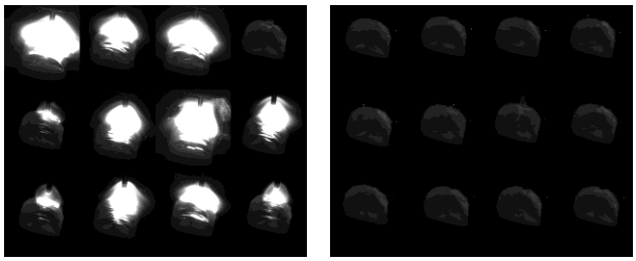
Figure 8. High frame rate molten pool image.

Figure 8 shows the weld pool images during pulse phase and CMT phase.Through high-speed camera, the droplet dripping and the process of drawing the wire can be clearly observed. The intensity of the CMT phase is significantly weaker than the pulse phase. Corresponding to the current data, it can be found that when the current has a slight rise, the wire contacts the weld pool, and the arc is extinguished and maintained for a period of time.

5. EXPERMENTAL RESULTS AND EXTENSIONS

5.1 Experimental results

The weld pool image acquisition experimental system is shown in Fig. 1, and monochrome camera is used to capture weld pool image. Part of the collected weld pool images are shown below.



(a) No trigger (b) Trigger

Figure 9. Monochrome weld pool image.

We define the overexposure ratio as the ratio of overexposed images in all images. If the total number of pixels with 255 is greater than 1% in the image, then it is defined as overexposed In the experiments, 500 weld pool images were captured and the overexposure ratio of all images was counted. With no trigger, most images have serious arc interference, and the image over-exposure ratio is higher than 95%, that is to say, most images cannot be applied to subsequent weld pool image processing. However, with trigger, the image over-exposure ratio is less than 3%, and there is basically no arc interference, which is more convenient for subsequent weld pool image processing.

5.2 Experimental extension

We put some asphalt on the 304 stainless steel and use the proposed system to capture weld pool images. Since the captured weld pool images have high SNR and almost no arc interference, therefore, the OTSU method is used to binarize the weld pool image and extract the width of the weld pool.

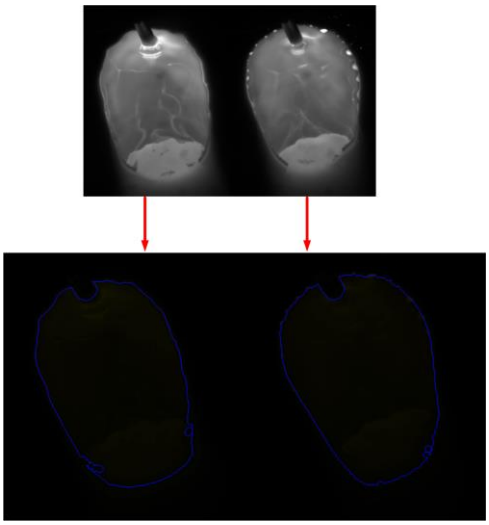


Figure 10. Contour change image.

As shown in Fig. 10, when passing through the asphalt region, since the proposed system completely eliminates arc interference, it is possible to clearly observe the width of the weld pool is significantly widened. The variation of the width of the weld pool throughout the welding process can be calculated as shown below.

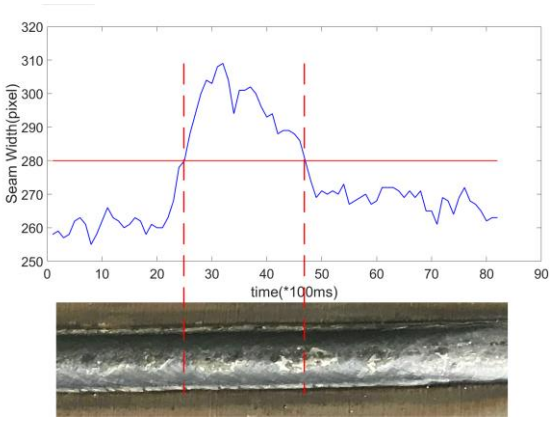


Figure 11. Width change image.

As shown in Fig. 11, the red dotted line area is the area with asphalt, and the weld pool image in the area is significantly widened.

6. CONCLUSION

(1) The FPGA was firstly used as the data acquisition and triggering synergy module of the weld pool image system, and then the data was processed with MATLAB. The combination of hardware and software is applied to the weld pool image acquisition system, which significantly improves the working efficiency of the whole system.

(2) Through using photodiodes and V611 high-speed digital cameras, the intrinsic relationship between light intensity and welding current is fully analyzed. Light intensity and welding current are combined to accurately determine the arc extinction time, thereby obtaining a weld pool image with high signal to noise ratio.

(3) Extensive online experiments were carried out to verify the accuracy and stability of proposed trigger system and realise the detecting change of weld pool width with asphalt.

7. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (no. 61727802 and 61501235).

8. REFERENCES

[1] Wang Ke-hong,You Qiu-rong, Hen Ying-ji. Preliminary discussion about image character of gas pore in MAG welding based on vision sensing[J]. Transactions of The China Welding Institution, 2006, 27(12):13-16.

[2] Tsai F R, Elijah Kannatey-Asibu J. Modeling of Conduction Mode Laser Welding Process For Feedback Control[J]. Journal of Manufacturing Science & Engineering, 2000, 122(3):420-428.

[3] Shi Y, Zhang G, Ma X J, et al. Laser-Vision-Based Measurement and Analysis of Weld Pool Oscillation Frequency in GTAW-P[J]. Welding Journal, 2015, 94(5):176-187.

[4] Scotti A, Morais C O, Vilarinho L O. The effect of out-of-phase pulsing on metal transfer in twin-wire GMA welding at high current level[J]. Welding Journal, 2006, 85(10): 225-230.

[5] Yan Zhi-hong,Zhang Guang-jun,Qiu Mei-zhen, et al. Monitoring and processing of weld pool images in pulsed gas metal arc welding[J]. Transactions of The China Welding Institution, 2005, 26(2):37-40.

[6] Zhang Hong-tao, Hong Ji-cai, Hu Yue-liang, Energy inpue and metal transfer behavior of CMT welding process[J]. Materials Science and Technology, 2012, 20(2):128-132.

[7] Li Meng-xing, Wu Yi-xiong, Cai Yan and Sun Da-wei, Research Review and Developmental Trends of Image Technology in Extracting Welding Pool Feature Parameters[J]. Hot Working Technology, 2010, 39(21):142-145.

[8] Motta, Dutra M F, Jr C G, et al. A Study on out-of-phase current pulses of the double wire MIG/MAG process with insulated potentials on coating applications: part II[J]. Journal of the Brazilian Society of Mechanical Sciences & Engineering, 2007, 29(2):207-210.

[9] Babkin A S, Gladkov E A. Identification of Welding Parameters for Quality Welds in GMAW[J]. Welding Journal, 2016, 95(1):37-46.

[10] Chen Yan-bing, Li Li-qun, Chen Feng-dong, et al. Application and prospect of image processing in welding[J]. Materials Science and Technology, 2003, 11(1):106-112.

[11] Nagesh D S, Datta G L. Prediction of weld bead geometry and penetration in shielded metal-arc welding using artificial neural networks[J]. Journal of Materials Processing Tech, 2002, 123(2):303-312.

[12] Clocksin W F, Bromley J S E, Davey P G, et al. An Implementation of Model-Based Visual Feedback for Robot Arc Welding of Thin Sheet Steel[J]. International Journal of Robotics Research, 1985, 4(4):13-26.

[13] Yong Z, Jiang L, Yunhua L I, et al. Welding Deviation Detection Algorithm Based on Extremum of Molten Pool Image Contour[J]. Chinese Journal of Mechanical Engineering, 2016, 29(1):74-83.

[14] Lin Shi-miao, Hign-speed signal acquisition based on FPGA[D]. Southwest Jiaotong University 2017.

[15] Zhang Dawei, Jiang Jing, Liu Di, Design of IIR Digital Filter Based on FPGA[J]. Marine Electric & Electronic Technology, 2012, 32(2):24-26.

Design a Real-Time Eye Tracker

Sara Bilal
Senior Lecturer
Whitieria New Zealand
+64211257123
Sara.Bilal@whitireia.ac.nz

Mohamad Hassrol Bin Mat Hussin
Engineer
IIUM
hassrol14@gmail.com

Rasheed Nassr
Senior Lecturer
University of Kuala Lumpur
rasheed@unikl.edu.my

ABSTRACT
An eye tracker is a device for measuring eye positions and eye movement. During the past decades, various approaches have been presented for eye tracking systems such as helping disabled people in various tasks including communication, writing emails, and drawing, besides other applications such as cognitive studies, electronic games and commercial eye trackers. The eye tracking devices and the algorithms that track the gaze still need an improvement. Many critical problems arise from the fact that the eye tracker system does not retain its accuracy for a long period of time. Besides, the fact that the commercial eye trackers are expensive. The aim of this research work is to develop a typing by gaze system using a PS3 camera. The developed typing by gaze system uses eyes' gaze for typing on a virtual keyboard. As well, the developed system is tested using a built-in laptop camera and a mounted PS3 camera on an eyeglass. First, the system will detect the face from the video stream. Followed by eyes' pupil detection and tracking which has been achieved using template-matching technique and adaptive EigenEye method. In addition, a virtual keyboard has been designed based on a laptop-size keyboard with QWERTY layout. After the eye pupil is tracked, the system will take control of the mouse's cursor movement. Then the eye's gaze will become the pointing device and the selection of the focused key or button will be achieved by using dwell-time of 0.5seconds. The developed eye tacker is an affordable device and it has retained a good typing accuracy either by using the PS3 or the laptop camera. Hence, it can be used for several applications and especially for research purposes.

CCS Concepts
•Computer systems organization ~ Real-time system architecture

Keywords
Eye tracking; Mounted camera; PS3 camera; Virtual keyboard

1. INTRODUCTION
In the recent years, eye-tracking systems turned out to be an important tool to increase the quality of life and enhance the autonomy of persons with disabilities needing alternative input devices [1][2]. An eye-tracker is a device that applies projection patterns and optical sensors to accumulate data about eye position, gaze direction or eye movements with very high accuracy. Most eye-trackers systems are actually based on the fundamental principle of corneal reflection tracking as shown in Figure 1.

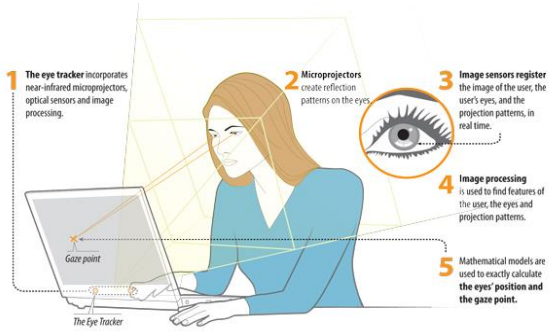


Figure 1: An eye-tracker [2]

A convenient, natural and high-bandwidth source of input is provided and can be attained by the movement of user's eye. Just by tracking the direction of the user's gaze, or the center pupil of the eye, the subject or information about what the user is looking at can increase the communication bandwidth from the user to the computer. However, to track the eye and its movement is not an easy task especially the pupil. Generally, eye movement is divided into two categories which are fixations and saccades. A fixation usually happens when the eye gaze pauses in a certain position. In the eye-tracking situation, fixations are used to denote a starting point for all eye movements. Saccades are when the eye gaze moves to another position. Humans alternate between fixations and saccades. This behaviour corresponds to the scan paths which are used to analyse cognitive intent, interest and importance. The fovea provides the bulk of visual information. The periphery is less informative than the fovea [3]. Figure 2 shows the fovea located in the macula region of the retina. The locations of the scan paths during eye-tracking signify the information that was processed. Fixations normally last for 200 ms when reading text and 350 ms when viewing a scene.

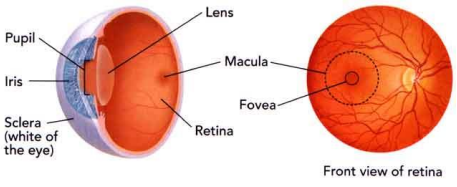


Figure 2: Details part of the eye [3]

In this paper, the existing eye-tracking systems has been surveyed and the methodology to design an affordable eye-tracker is described (see Figure 3). The idea of the system is to mount an infrared camera onto a pair of eyeglass and capture the movement of the pupil using a software. Following that, the developed typing by gaze system was conducted using two systems: 1) a distanced

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301509

camera which requires the detection of the whole face and then locate the eye. 2) A wearable camera device by focusing the camera on the eye.

2. RESEARCH BACKGROUND

2.1 GAZE INPUT

Gaze is naturally used to obtain visual information. For example, gaze location shows the focus of attention. As an input method, gaze has both advantages and disadvantages. It is a natural mode of input as it is easy to focus on items by looking at them. Another advantage is that the target acquisition using gaze is very fast, provided the targets are large enough. However, gaze is not as accurate as the computer mouse. Inaccuracy originates partly from technological reasons and partly from features of the eye [4]. The size of the fovea and the inability of the remote camera to resolve the fovea position restrict the accuracy of the measured point of gaze to about 0.5 degrees, equivalent to a region spanning approximately 15 pixels on a typical display (17 inch display with a resolution of 1,024 × 768 pixels viewed from a distance of 70 cm). One problem is drifting; even if a newly calibrated eye tracking device is accurate at first, with continued use the measured point of gaze drifts away from the actual point of gaze. This is partly due to the technology and partly due to the interaction between head movement and eye movement. Therefore, the practical accuracy is about 0.5–1 degree of drifting corresponds to about 1–1.5 cm on the screen at a normal viewing distance. When we look at things, we fixate (focus) on them, with fixations typically lasting from 200 to 600 ms [4]. For a computer to distinguish whether the user is looking at an object to obtain information or to select it, an interval longer than the typical fixation interval is needed. Stampe and Reingold [5] used a dwell time (an extended look at the object) of 750 ms in their eye typing study. A thousand milliseconds is usually long enough to prevent false selections. For simple tasks, 700 ms or less is enough. Requiring the user to fixate for long intervals is good for preventing false selections (thus, preventing the Midas touch problem), but this is uncomfortable for most users [5]. It is important to note that the use of a dwell time criterion for key selection places an upper limit on eye typing speed. In other words, these is no skilled acquisition system exists that will allow a user to "eye press keys" at a rate faster than 1 / td, where td is the dwell time. If, for example, td = 1,000 ms = 1 s, the upper limit for typing speed is (60 / 1) / 5 = 12 words per minute (wpm) (following the accepted method in computing typing speed of 1 word = 5 inputted characters).

2.2 COMPARISON BETWEEN the TYPING BY GAZE SYSTEMS

Nowadays, there are many systems of typing by gaze such as Tobii PCEye Go [6], The Eye Tribe Tracker [7], Eyegaze Edge Desktop [8] and myGaze Assistive 2 [9]. The Tobii PCEye Go is a peripheral eye tracker that enhances computer accessibility with the speed, power and accuracy of gaze interaction. The device replaces the standard mouse, allowing you to navigate and control a desktop or laptop computer using only your eyes. Tobii PCEye Go system is using an integrated keyboard or built in keyboard to write texts or to enter a web address. Whereas for The Eye Tribe Tracker, the information extracted from person's face and eyes is used to calculate the location from where a person is looking. The calculation is based on a pair of eye gaze coordinates represented by (x, y) on the screen coordinate system. The Eyegaze Edge Desktop is an eye-operated communication and control system that empowers people with disabilities to communicate and interact with the world. By looking at control keys or cells displayed on a screen, a user can generate speech either by typing a message or selecting pre-programmed phrases. Eyegaze edge systems are used to write books, attend school and enhance the quality of life of people with disabilities all over the world. This system has an adjustable monitor arm with camera bracket and using a high-speed infrared sensitive camera and lens. Eyegaze Edge Desktop system already design their own on-screen keyboard for the typing. In addition, MyGaze Assistive 2 is a gaze control system using computer programs which is affordable. It is well suited for special needs students and their teachers as well as adults who rely on gaze for communication and environment control. The myGaze Eye Tracker relies on a technology by SensoMotoric Instruments (SMI) from Germany, a leading developer of state-of-the-art eye tracking solutions for more than 20 years. This system has less than 50ms system latency, strong robustness and reliable performance. MyGaze Assistive 2 is also an Instant gaze tracking that its calibration is minor. The system will instantly tracks the users gaze. The size of these products is extremely portable, except for the Eyegaze Edge Desktop which is bigger like laptop-size. Table 1 shows a clear view of the comparison however, there are some factors that affect the commercial product such as cost effectiveness, robustness under different environments, real-time application speed, and end-user acceptability.

Table 1: Comparison between existing typing by gaze systems

system	Camera	keyboard	Price	Advantage(s)	Limitations
Tobii PCEye Go [6]	Infrared LED	On-screen keyboard integrated	$1,995	- user can wear contact lenses. -Can hit the smallest targets	Expensive
The Eye Tribe Tracker [7]	Infrared LED	On-screen keyboard	$99	-Very cheap -USB 3.0 superspeed	user can't wear contact lenses
Eyegaze Edge Desktop [8]	Infrared mounted camera	On-screen keyboard	£4,250-$6,446.21	Fastest eye gaze system on the market	-Very expensive -Bigger
myGaze Assistive 2 [9]	Infrared LED	On-screen keyboard	€499-$557.83	-Cheap -Calibration-less mode	Hard to pause interaction

On the other hand, there are several existing free eye tracker software such as IMOTIONS, XLAB [10]. However, their cons is more than their pros. Hence, developing an affordable eye tracker device apart from the computer cost is a demand in the research sector.

3. METHODOLOGY

3.1Eye Tracking Systems

Developing an eye tracker can be based on tracking the eye from a distance or using a wearable camera as illustrated in Figures 3(a) & (b).

(a) **(b)**

Figure 3. (a) Distanced camera [6], (b) Wearable Camera device.

3.1.1 Eye Tracker Using a Wearable Camera

The process of acquiring eye image is done using a PS3 camera that is mounted on a eyeglass. To build this device, the components are listed in Table 2 with their respective price in Ringgit Malaysia (*RM*).

Table 2: Cost summary to build an eye-tracker

No	Components	Quantity	Price per unit (*RM*)	Price (*RM*)
1	PS3 Eye Camera	1	88.00	88.00
2	Eyeglass	1	15.00	15.00
3	IR LED Transmitter	1	1.50	1.50
4	IR LED Receiver	1	1.50	1.50
5	Alligator clips	4	2.60	10.40
6	Pack of wire ties (short)	1	1.50	1.50
7	Pack of wire ties (long)	1	1.90	1.90
8	AAA size battery holder	1	1.00	1.00
9	3 × AAA battery holder	1	1.50	1.50
10	4 × AAA battery holder	1	2.00	2.00
11	9mm Aluminium Wire Gauge	1	15.00	15.00
12	Camera Lens Mount	1	14.00	14.00
13	IR infrared filter gel	1	20.00	20.00
14	Electrical junction connectors	2	13.00	26.00
15	Screws	5	2.20	11.00
			Total	210.30

3.1.1.1. PS3 Camera Configuration

PS3 camera is chosen among all other cameras because it small in size, performs well in variable lighting conditions, has stable USB performance, and capable of high frame rates. It also comes with the resolution of 640×480 pixels which is at 60Hz frequency that gives a solid performance. Also, an eye image that captured by this PS3 camera will give a smoother image since the maximum video frame rate is 120fps. The higher the frame rate, the smoother the screenplay and the image capturing. From the optics perspective, it uses wide angle lens with field of view of 75°. Moreover, the minimum focus distance allowed are 9.84 inches ≈ 25 cm, and the low light adjustment feature also available. PS3 camera is used because even when there is less light, the PS3 camera's infrared light can still provide tracking information with high accuracy because of the presence of the IR light. Basically, the PS3 camera will be dissembled from its body frame and mounted on the eyeglass and the eyeglass's lens will be removed from its frame too. The wire tier is used to secure and tie the camera arm to the eyeglass's frame.

3.1.1.2 Capturing Eye Image Using Wearable Camera

To capture an eye image, the camera is mounted on the eyeglass and is located right in front of the patient's eye so that it can capture the eye image accurately. The camera is selected based on the quality images captured, the frequency and price. There are some cameras which have been used for eye-tracking such as HATCAM, basic web camera and PS3 camera [11]. The distance between the camera and eye must be seriously considered. The distance cannot be too far or too close to the eye. Hence, in this project, the distance of the camera to the eye is fixed to 3.2 cm.

3.1.2 Eye Pupil Detection And Tracking

Most of the image based methods are detecting the eyes' location using features of the eyes which are known as knowledge-based methods, feature-based methods, simple template-matching and appearance-based methods. Another interesting method is "Deformable template-matching" which is based on matching a geometrical eye template on an eye image by minimizing the energy of the geometrical model. To locate and track the eye pupil, the original image of the eye must be converted to greyscale image first. Then, we need to apply two methods that is template matching technique and adaptive EigenEye method.

3.1.2.1 Eye Pupil Tracking Using Adaptive EigenEye Method

EigenEyes is used as core components of the original images. This EigenEyes is the product of decomposition of eyes image by the adaptive EigenEye method. These Eigeneyes have a role as the orthogonal basis vectors of a subspace called eyespace. In this work, we use the Principal Component Analysis (PCA) methods to extract features for eye tracking. PCA is a well-known unsupervised algorithm for linear feature extraction; it is a linear mapping that uses the eigenvectors with the largest eigenvalues. The advantage from this method is the detection process is a simple step throughout the whole algorithm. For PCA, a set of 860 right eye images, 860 left eye images and 320 non eye images have been used. The eye-template image of 40×26 pixels in the training database is defined as the column vector of 1040×1 pixels. The column vector is then reduced to an eigenvector with dimensions by using PCA, which seeks a projection that best represents the original data in a least-squares sense.

3.1.2.2 The localisation of eye centers

There are many commercial products to locate and track the eye gaze such as [6, 7] with are expensive hardware headset. In this research work, webcam gaze tracker which is based on a simple image gradient-based eye center algorithm is used. Timm and Barth, (2011) proposed image gradients-based approach to accurately find the eye centre. Their approach was used here because they have used simple objective function, which consists of dot products. The maximum of this function matches the location where most gradient vectors intersect and thus to the eye's centre. This simple approach, it is invariant to changes in scale, pose, contrast and variations in illumination [12]. Therefore, in an environment with a low resolution images and real time application this approach is good enough to be used for gaze tracking. To determine the eye region fractions, let (x, y) be the upper left corner and W, H the width and height of the detected face. Then, the mean of the right eye centre is located at (x + 0.3, y + 0) and the mean of the left centre is at position (x + 0.7, y + 0.4). The Gaussian blur was used before processing face to smooth noise. The gradient algorithm used in this implementation is as follows: The gradient function creates vectors that always point towards the lighter region. Since the iris is darker than the sclera, the vectors of the iris edge always point out. It means that at the center they will be facing in the same direction as the d vector. This method has simplified tracking a reference point such as eye corner to accurately judge where the user is looking. However; this approach first detect the face then determine the location of the eye's center later; this approach wouldn't help with application that require using camera that is only focused on one eye in order to accurately track eye movement. Even though the previous mentioned approach works, however tracking eyes by using a special camera focused on one eye seems to be more efficient tracking the eyes.

3.1.2.3 Track Single-Eye Pupil Using PS3 Camera

To track an eye and find the center coordinates of pupil, Houghcircles was excluded because it didn't perform well and hence, hue, saturation and val (HSV) filter was used. First, the images are thresholded then a contour finding algorithm was used to find the centroid of all the contours. This gives the center coordinates of the eye pupil. This method was working with high efficiency in real time even when the eye is blinking. In this work, PS3 camera was used. This camera comes with API that can help to run this camera. Then by integrating the previous code with the API of the PS3 camera, we develop a prototype using VC++ 2010 with OpenCV. Sample pictures generated from previous code are presented in the Figure 4.

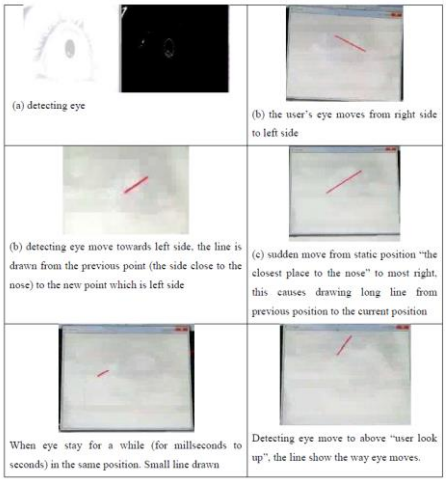


Figure 4: Tracking eye results

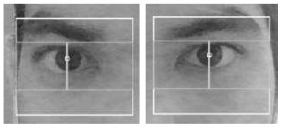
4. Experimental Results

4.1 Eye Pupil Detection and Tracking

First, the camera captured the image or the frames that contain the face and eye pupil. Following that, the image is converted to grayscale and the face is detected by using Haar-like features [13]. Figure 5 shows the result of face detection. The blue rectangle indicates that the face region has been successfully detected.

Figure 5: Face detection.

Next, the captured image also contains the eye pupil. Template-matching technique and EigenEye method has been used to track the eye pupil as shown in Figure 6.



(a) (b)

Figure 6: Pupil detection of (a) Right eye and (b) Left eye.

4.2 Mouse Control

After the tracking process is done, the system will proceed to the mouse control. It will capture the current position of mouse first. Then by using some calculation it will set the position of mouse cursor corresponding to the focus point. The movement of eye will move the mouse cursor. Figure 7 shows the result of scaling up the mouse pointer. Because of the small size of eye pupil region compared to the resolution of desktop, the scale factor or ratio become too big. Thus, it make '1' pixel of eye pupil movement equal to about '45' pixels horizontal movement and '41' pixels vertical movement on the actual screen. Therefore, it will reduce the accuracy of the cursor as the buttons size is about '70' pixels square.

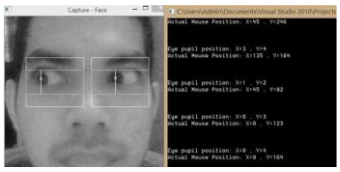


Figure 7: Result of the mouse position around the initial point

4.3 Virtual Keyboard

4.3.1. Virtual Keyboard Design

The virtual keyboard was designed using Visual Basic 2010. The keyboard is using standard-laptop keyboard and QWERTY layout. (See Figure 8)

Figure 8: Virtual keyboard for this gaze typing system.

4.3.1.1 Virtual Keyboard Layout

For the virtual keyboard, when any button is pressed, the system will copy the name or the text of that button and send into textbox. For example, when the user clicks button 'e', the word 'e' will be sent to the textbox (see Figure 9 (a)). If the person needs to use any symbol that is not active at that time which means behind the numeric button, he/she needs to use the "SHIFT" key to activate all the symbols and presses or clicks the desire symbol (see Figure 9(b)). The user also can copy any of the word or sentence by pressing the "COPY" button and paste back to textbox by pressing "PASTE" button (see Figure 9 (c)). If user want to clear everything in the textbox can press "CLEAR" button which will clear everything automatically rather than using "BACKSPACE" button repeatedly (see Figure 9 (d)).

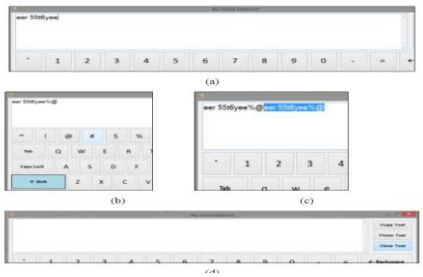


Figure 9: (a) Some buttons are clicked; (b) SHIFT button is clicked to active the symbols and capital letter; (c) COPY and PASTE button are clicked; (d) Textbox is cleared

4.4 Dwell-time Method

Dwell-time method means that the mouse will automatically be clicked when the focus time has reached the desired limit which is 0.5 second. Focus time will start counting when the mouse pointer enters or hovers to any button. The focus time will start new count when the previous count has reached 0.5 second. This developed system also incorporates a sound that will play automatically on the same time with the click. This can increase the typing speed without having to look at the textbox to see whether the character is typed or not because the sound has indicated that the typing is successfully performed.

4.5 FINAL RESULTS

The pupil was extracted to exploit its higher contrast than the background due to the IR components of the natural light. Segmentation is applied to detect the pupil successfully using grey level images and other features' extraction techniques. Following that, the system took control of the mouse cursor. So the users can use their eyes for typing by clicking the buttons on the virtual keyboard. This is achieved by implementing the dwell-time method. The time for an automatic click is 500ms (0.5 second). So the user has to focus on any button at the virtual keyboard. Then the button will be highlighted and automatically send the button's text into textbox or message box. (See Figure 10)

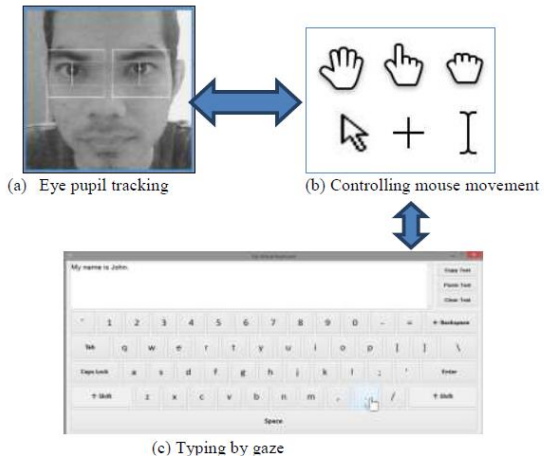


Figure 10: Process of typing using the eye gaze

5. CONCLUSION

Eye tracker is a communication and control system that has been implemented in many applications and it empowers people with disabilities (parallelized people) to communicate and interact with the environment. There are two methods to develop an eye tracker which are based on tracking the eye from a distance or by using a wearable camera. However, by using a wearable camera, the eye tracking system can perform better. The existing eye-trackers are quite costly, therefore the cost of building an eye-tracker using PS3 is affordable. After constructing the eye tracker hardware where an infrared camera has to be mounted onto an eyeglasses and capture the movement of the pupil using VC++ and OpenCV. Following that, the template matching method is used to locate and track the centre of the pupil, where a small patch of dark pixels are used as a template. The search space in each frame could be defined based on the estimates of the pupil location from the previous frames. As a result, this affordable constructed eye tracker can be used to conduct research as well it can help disabled patients in various tasks such as communication, writing emails, drawing and gives them a sense of interaction with the surroundings. In the future, a mobile eye tracker device can be developed using these approaches.

6. REFERENCES

[1] Alan Watt, Fabio Policarp, The Computer Image, Addison Wesley, (1999). Antonio Lanatà,

[2] Lemahieu , W., & Wyns, B. (2011). Low cost eye tracking for human-machine interfacing. Journal of Eye Tracking, Visual Cognition and Emotion nº (8), http://hdl.handle.net/10437/22971

[3] "Age-Related Macular Degeneration" Retrieved March 13, 2016 from http://www.urbaneyemd.com/glossary_details.php?id=77

[4] Jacob, Robert JK. 1995. Eye tracking in advanced interface design. In: Barfield W, Furness TA (eds) Virtual environments and advanced interface design. Oxford University Press, New York, pp 258–288.

[5] Ware C., Mikaelian H.H. 1987. An evaluation of an eye tracker as a device for computer input. In: Proceedings of CHI/GI '87. ACMPress, Cambridge, pp 183–188. DOI 10.1145/29933.275627

[6] "What is eye-tracking". Retrieved April 4, 2014 from http://www.tobii.com/en/about/what-iseye- tracking/

[7] TheEyeTribe. 2015. The eye tribe tracker. Retrieved April 4, 2014 from https://theeyetribe.com/products/

[8] XLab.. 2015. Xlabs, eye gaze and head via the webcam. Retrieved April 4, 2014 from http://xlabsgaze.com/

[9] MyGaze Eye Tracker, Retrieved April 4, 2014 from http://www.mygaze.com/products/mygaze-eye-tracker/

[10] 10 Free Eye Tracking Software Programs. Retrieved 5/2/2018 from https://imotions.com/blog/free-eye-tracking-software/,

[11] Takayuki Ito, Kyle Mcdonald, Golan Levin and students . 2010. Eyewriter collab at Parsons MFADT, (2010). The Eye-Writer 2.0.

[12] Timm, F., & Barth, E. 2011. Accurate Eye Centre Localisation by Means of Gradients. In VISAPP (pp. 125-130). https://doi.org/10.5220/0003326101250130

[13] Viola, P., & Jones, M. J. 2004. Robust real-time face detection. International journal of computer vision, 57(2), 137-
54. https://doi.org/10.1023/b:visi.0000013087.49260.fb

Using Bezier Curves to Refine Road Vector Data through Satellite Images

Maneesha Perera
University of Colombo School of Computing
Colombo, Sri Lanka
(+94) 774952102
maneeshagsp93@gmail.com

Damith Karunaratne
University of Colombo School of Computing
Colombo, Sri Lanka
(+94) 718581379
ddk@ucsc.cmb.ac.lk

Enosha Hettiarachchi
University of Colombo School of Computing
Colombo, Sri Lanka
(+94) 714233694
eno@ucsc.cmb.ac.lk

ABSTRACT
Availability of various geospatial data has increased with the recent growth of geospatial information on the web. Integrating or conflating multiple geospatial datasets have become an important aspect in modern geographic information processing as it can provide insights that are not capable to obtain from an individual dataset. However, positional inconsistencies or misalignments can occur during the conflation process as different geospatial datasets could have different accuracy levels, different geospatial data types (vector or raster), and different projections when they originate from different sources. Thus, making the conflation process a challenging task. In this paper, we present a novel alignment algorithm which uses the concept of Bézier curves to refine the positional inconsistencies that occur during the vector to imagery conflation process of road vector data. A novel evaluation model which considers the area of the polygon bordering the actual road and the vector layer is introduced. Proposed approach is evaluated against the existing Piecewise Linear Rubber Sheeting method on the Sri Lankan road vector data on test scenarios consisting of road segments with varied misalignments. Results show that the proposed bézier approach has an average misalignment reduction percentage of 75.8% compared to 68.6% from the existing piecewise linear rubber-sheeting method.

CCS Concepts
• **Information systems ~ Geographic information systems**
• **Computing methodologies ~ Computer graphics**

Keywords
Bézier; Conflation; Raster; Vector; GIS; Satellite Images;

1. INTRODUCTION
Geographic Information System (GIS) is a computer-based tool for mapping spatial data. It encompasses functionality for querying and statistical analysis of data as any other common database management systems while providing data visualization and geometry analysis features unique to geographic data. Data in GIS represent an abstraction of physical entities such as roads, mountains, accident locations or other features which interest the user.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301512

Two basic types of data models known as vector data and raster data (e.g. satellite images, aerial images, scanned maps) are utilised for storing spatial data digitally in GIS systems. Raster data models represent the world as a set of cells in a grid pattern, where each cell typically represents an area on the earth surface. Vector data models use a set of coordinates and associated attribute data to define discrete objects. Coordinates define spatial locations and shape, attributes record the important non-spatial characteristics. Important data of base maps are generally stored in raster data. But most of the manipulations are carried out using vector data models. Due to the availability of various geospatial in the current context, integrating multiple geospatial datasets has become an important aspect in the study of GIS [3,8,12,13].

In GIS research, *conflation* is the often-used term for integration of multiple geospatial data from different sources [10]. Specifically, integration or conflation is the process of combining two or more spatial representations of the same region to produce a superior dataset better than any of the original datasets. Through conflation, individual strengths of different datasets can be aggregated. However, since these geospatial datasets originate from different sources positional inconsistencies can occur and they often do not match well to each other [3,10,12]. Further, since this positional inconsistency is nonsystematic, carrying out a simple global transformation will not solve this problem. Manual correction of these inconsistencies can be very repetitive, labor intensive, and time consuming which is often not practical [12]. Thus, conflation is a challenging process. Figure 1 shows a real-world example of these positional inconsistencies (misalignments or deviations) between the existing road vector data of the Sri Lanka survey department and the road layer represented by the satellite image of Google maps in a scale of 1:10,000. Black lines represent the vector layer and the white lines in the satellite image represents the actual road layer. The arrows shown in yellow showcase the deviation of the vector layer from the actual road segment.

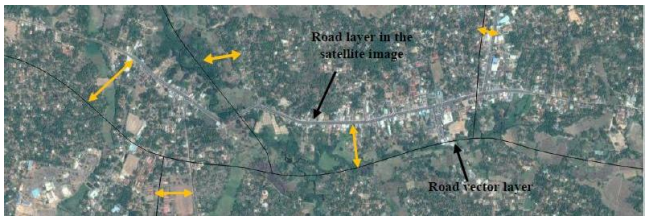


Figure 1. Misalignment between the road vector data and satellite image.

The rest of the paper is structured as follows. Section 2 outlines the related work on conflation approaches, Section 3 describes the

proposed research methodology, Section 4 elaborates on the evaluation model and the results obtained for the proposed approach, and Section 5 provides a conclusion and outlines the future work.

2. RELATED WORK

Chen et al. [3] classified the conflation techniques based on the geospatial data types into three categories as vector to vector data conflation (e.g. integration of two road networks of different accuracy levels.), vector to raster data conflation (e.g. integration of road network and imagery.) and raster to raster data conflation (e.g. integration of raster maps and imagery.).

Having an accurate vector geographic database is an important aspect in GIS [2,3,10]. Thus, many GIS researches have been focused on vector to vector conflation in the past few decades [1,13,15]. However, due to the advancement of the remote sensing technology to capture high resolution imagery, vector to imagery conflation has gained importance [3,4,8,9,12], as this accurate imagery can be used to generate a precise vector geographic database.

Vector to imagery conflation approaches consist of two main steps. Finding accurate control point pairs and applying a conflation algorithm making use of the identified control point pairs.

2.1 Extracting Control Points

To perform vector to imagery conflation, some spatial objects from the imagery should be extracted to serve as control points. Control points can be considered as counterpart elements of the vector dataset and imagery. These are the base points used to carry-out the refinement process. The existing vector to imagery conflation approaches have utilized different methods to extract these counterpart elements. Various GIS researchers and computer vision researchers have shown that, the intersection points on the road networks are good candidates to be identified as an accurate set of control points, as road intersections are salient points to capture the major layouts of the road network and the road shapes around the road intersections are well defined [5]. Thus, several approaches have been carried out to identify road intersections from the imagery and vector dataset for the conflation process. Chen et al. [4] carried out a localized image processing technique to identify road intersections. However, road segments in the satellite image near the intersection point of the vector data should be extracted in this technique which is a computationally intensive task. This approach was modified by Chen et al. [3] with the introduction of a histogram-based classifier to more accurately identify road intersections as control points and improved the localized image processing technique by exploiting road vector direction and widths to generate templates to match against the satellite imagery. This approach can only refine vector layer misalignments within a limited range as the run time increases when the search area size increases for the template matching. Another main issue with template matching approaches is that only limited shape models can be provided. It is possible for the angle and road branch width to vary significantly even for one type of intersections in an image, and this may cause matching issues. Also, partial occlusion near intersection points will cause problems for the matching process.

2.2 Refinement Using a Conflation Algorithm

The second and most important step of the vector to imagery conflation techniques is the refinement process using a conflation algorithm. However, only two main algorithms such as Piecewise Linear Rubber-Sheeting algorithm and Snakes algorithm have been primarily used. From these two algorithms the Piecewise Linear Rubber-Sheeting algorithm plays an important role in vector to imagery conflation approaches.

The piecewise linear rubber-sheeting technique for map transformation introduced by White and Griffin [16] has been utilized by many researchers to align the vector dataset with the imagery after the control point identification process [3,4,12,14]. The algorithm assumes the map to transformed as a rubber-sheet which is stretched to coincide with the stable map. This two-step approach starts with the triangulation process which divides the rubber-sheet map into triangles based on the provided control points. It is important to avoid long and narrow triangles. Thus, a quadrilateral test is carried out to ensure the triangulation maximizes the minimum triangle height. The second step is the transformation process which maps the individual triangles in the rubber-sheet map onto its corresponding triangle on the stable map using only an affine transformation. Piecewise rubber-sheeting method based on triangles with extremely small angles (long and thin triangles) results in contorted conflation. Thus, Chen et al. [4] and Chen et al. [3] performed Delaunay triangulation [6] to avoid triangles with extremely small angles. In small regions of the selected control points, the results of the rubber-sheeting approach will be quite good, because it forces those points to coincide, but at places far from the control points, misalignments may remain. These misalignments can also be refined by adding more control points with the cost of making the process more complicated. Further, this approach is limited in coping with distortions between sets, since only the seed points are matched and the relative disagreements between the linear features shapes remain unresolved. However, since this approach is considered with generating triangles based on the control points, it is necessary to have a sufficient number of control points to generate the triangles. The computation cost of the algorithm is associated with the triangulation and transformation processes. The triangulation process has $O(nlogn)$ worst case time complexity, where n is the number of control points.

However, as stated above these the conflation algorithms consist of both advantages and disadvantages that directly impact the vector to imagery conflation process. Therefore, in this paper we propose a conflation algorithm which addresses aforementioned limitations.

3. METHODOLOGY

The research methodology includes three main steps: Coordinate Reference System (CRS) Transformation, Region Selection, and Refinement. Figure 2 represents a high-level view of the proposed methodology. Initially, the existing road vector layer database will be decomposed into a set of points (a set of points will represent each road segment in the vector database). Therefore, transforming these points will result in a transformation of the road layer. Secondly, these data points and the satellite image will be transformed to a single CRS as this transformation is important to see the misalignments between the two geospatial datasets. The focus is to carry out a local refinement for each region. Thus, a segmented region will be considered for the refinement process which contain several important steps. The first step would be the identification of a suitable scale to start the refinement. Scale identification is an essential aspect of the refinement process, as when the scale reduces the misalignment of the vector layer can be clearly visible. After identification of the initial scale, a set of control points from the vector layer and the satellite image will be provided. The control points from the vector layer and the satellite image can be junction points or road termination (end) points. Based on the provided control points the refinement for the road vector layer can be carried out considering the concept of Bézier

curves. The final output after the refinement will be an updated road vector database.

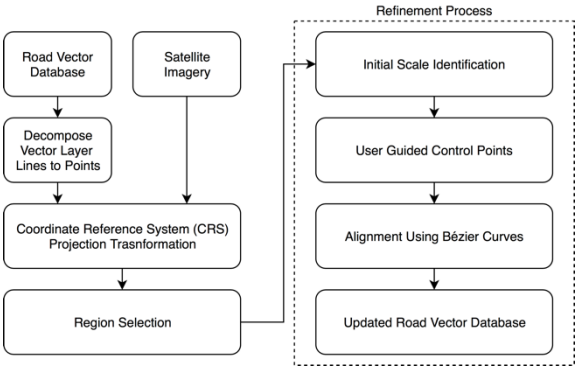


Figure 2. High level diagram of the methodology.

3.1 Alignment with Bézier Curves

The proposed solution for the refinement process of the road vector layer is alignment of the road vector considering Bézier curves. Bézier curves are parametric curves generated using linear interpolation, which are mostly used in computer graphics and related fields [7,11]. A Bézier curve is defined by a set of control points 0 to n and the number of control points define the order of the curve ($n = 1$ for a linear curve, $n = 2$ for quadratic curve etc.). Thus, an order n bézier curve is defined by $n + 1$ control points. The first and the last control points represent the end points of the curve while the other intermediate control points contribute to shape of the curve. Also, the intermediate control points do not lie on the curve.The formula for an order n bézier curve can be expressed by Equation 1.

$$B(t) = \sum_{i=0}^{n} \binom{n}{i} (1-t)^{(n-i)} t^i P_i \qquad (1)$$

Here $\binom{n}{i}$ are the binomial coefficients, P_i are the control points, and $0 \leq t \leq 1$. In this approach, we consider a road segment to be represented as an order two or order three bézier curve. Thus, assuming a road segment to be controlled by three or four bézier control points.

3.1.1 Alignment with Order Two Bézier Curves

The refinement of the vector layer considering order two bézier curves is carried out as described below. First, consider $V_1(x_1, y_1)$, $V_2(x_2, y_2)$ to be the provided vector layer control points and $R_1(x'_1, y'_1)$, $R_2(x'_2, y'_2)$ as the corresponding control points in the satellite image of the road segment to be refined (as shown in Figure 3).

Figure 3. Control points of the road in the satellite image (R1,R2) and control points of the vector layer (V1,V2)

Identify the points P_0, P_2 corresponding to $t = 0$ and $t = 1$ which are the start (V_1) and end points (V_2) of the road segment (These two points will act as the start and end control points for the bézier curve). Then, assign t values to the intermediate points in the road segment assuming that points are uniformly distributed. ($0 < t < 1$). Considering the road segment to be a quadratic bézier curve, equation 2 will represent the formula for a road segment.

$$B(t) = (1-t)^2 P_0 + 2t(1-t)P_1 + t^2 P_2 \qquad (2)$$

Next, find the remaining bézier control point P_1(as shown in Figure 4a). Coordinates of P_0 and P_2 are already known from step two as these are the vector layer control points. Therefore, from equation 2 we can formulate equation 3.

$$P_1 = \frac{B(t) - (1-t)^2 P_0 - t^2 P_2}{2t(1-t)} \qquad (3)$$

Consider any situation when $t = l$ and $B(l) = (l_x, l_y)$. From step 3 we have already assigned t values for the vector layer points. Therefore, $B(l) = (l_x, l_y)$ is already known (l_x is the x coordinate of the vector layer point when $t = l$ and l_y is the y coordinate of the vector layer point when $t = l$). Thus, x and y coordinates P_{1x} and P_{1y} of control point P_1 can be derived using Equation 4 (similar equation for y coordinates). Where P_{0x} , P_{0y} are the x and y coordinates of point P_0 . Similarly, P_{2x} , P_{2y} are the x and y coordinates of point P_2.

$$P_{1x} = \frac{l_x - (1-l)^2 P_{0x} - l^2 P_{2x}}{2l(1-l)} \qquad (4)$$

Find the new location P'_1 of P_1 considering relative distance, if P_0 is mapped to R_1 and P_2 is mapped to R_2 (Figure 4b). This transformation will yield to transforming the road vector layer points. Finally, find the new coordinates of the vector layer points considering the assigned t values in step 3 and P'_1, R_1, R_2 and applying to Equation 5, which represents the transformed curve (road segment) as shown in Figure 4b.

$$B(t) = (1-t)^2 R_1 + 2t(1-t)P'_1 + t^2 R_2 \quad (5)$$

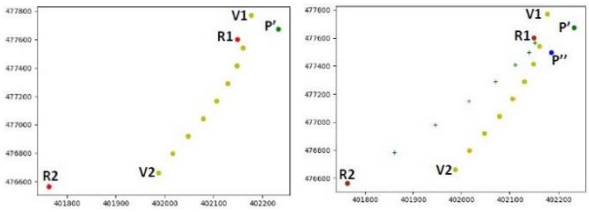


Figure 4. (a) left (b) right - Refinement Process for Order Two Bézier Curves

3.1.2 Alignment with Order Three Bézier Curves

The refinement of the vector layer considering order three bézier curves is carried as described below. First, consider $V_1(x_1, y_1)$, $V_2(x_2, y_2)$ the provided vector layer control points and $R_1(x'_1, y'_1)$, $R_2(x'_2, y'_2)$ as the corresponding control points in the satellite image of the road segment to be refined (as shown in Figure 3).

Identify the points P_0, P_3 corresponding to $t = 0$ and $t = 1$ which are the start (V_1) and end points (V_2) of the road segment (These two points will act as the start and end control points for the bézier curve). Then, assign t values to the intermediate points in the road

segment assuming that points are uniformly distributed. $(0 < t < 1)$. Considering the road segment to be a cubic bézier curve, equation 6 will represent the formula for a road segment.

$$B(t) = (1-t)^3 P_0 + 3t(1-t)^2 P_1 + 3t^2(1-t)P_2 + t^3 P_3 \quad (6)$$

Next, find the remaining bézier control points P_1, P_2 (as shown in Figure 5b). Coordinates of P_0 and P_3 are already known from step 2 as these are the vector layer control points). Consider two situations when $t = t_1, t = t_2$, and $B(t_1) = (t_{1x}, t_{1y})$, $B(t_2) = (t_{2x}, t_{2y})$. From step 3, we have already assigned t values for the vector layer points. Therefore, $B(t_1)$ and $B(t_2)$ is already known. (t_{1x}, t_{2x} are the x coordinates of the vector layer points when $t = t_1$ and $t = t_2$ respectively and t_{1y}, t_{2y} are the y coordinates of the vector layer points when $t = t_1$ and $t = t_2$ respectively). Thus, x and y coordinates P_{1x} and P_{1y} of control point P_1 and x and y coordinates P_{2x} and P_{2y} of control point P_2 can be derived by Equation 7 (similar equation for y coordinates). Where (P_{0x}, P_{0y}) and (P_{3x}, P_{3y}) are the x and y coordinates of point P_0 and P_3.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ (1-t_1)^3 & 3t_1(1-t_1)^2 & 3t_1^2(1-t_1) & t_1^3 \\ (1-t_2)^3 & 3t_2(1-t_2)^2 & 3t_2^2(1-t_2) & t_2^3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} P_{0x} \\ P_{1x} \\ P_{2x} \\ P_{3x} \end{bmatrix} = \begin{bmatrix} P_{0x} \\ t_{1x} \\ t_{2x} \\ P_{3x} \end{bmatrix} \quad (7)$$

Find the new location P_1', P_2' of P_1, P_2 considering relative distance, if P_0 is mapped to R_1 and P_3 is mapped to R_2 (Figure 5c). This transformation will yield to transforming the road vector layer points. Finally, find the new coordinates of the vector layer points considering the assigned t values in step 3 and P_1', P_2', R_1, R_2 and applying to Equation 8, which represent the transformed curve (road segment) as shown in Figure 5d.

$$B(t) = (1-t)^3 R_1 + 3t(1-t)^2 P_1' + 3t^2(1-t)P_2' + t^3 R_2 \quad (8)$$

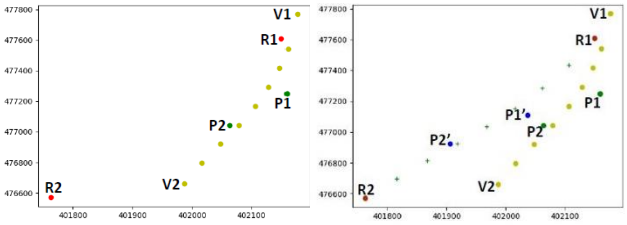


Figure 5. (a) left (b) right - Refinement Process for Order Three Bézier Curves

3.2 Evaluation Method

To evaluate the accuracy of the proposed refinement approaches the area of the polygon bordering the real road layer in the satellite image and the road vector layer is considered as illustrated in Figure 6. The proposed solutions were implemented using python 2.7, QGIS software, and PostgreSQL database management system with PostGIS (a spatial database extender for PostgreSQL). The road vector layer database of the Sri Lankan Survey Department and the satellite imagery of Google maps was utilized for evaluation. The real road segments of the satellite image were manually marked to prepare the ground truth data. The satellite image and the vector data were transformed to the Sri Lankan CRS. An image scale of 1:3000 was considered to identify the misalignments between the road vector data and the road layer in the satellite image.

The proposed solutions were tested on four scenarios which consists of different road misalignments. Test scenarios 1, 2, 3, 4

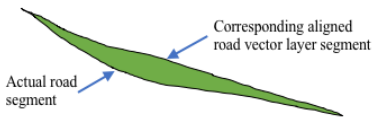


Figure 6. Polygon generated between the actual road and the aligned road.

consists of 5, 7, 10 and 9 misaligned road segments respectively. Test scenarios 1 and 2 contains simple road segments with one or two bends while test scenarios 3 and 4 contains complex road segments with multiple bending points. The piecewise linear rubber-sheeting algorithm was implemented, and the results were compared with the proposed approaches using the proposed Misalignment Reduction Percentage (MRP) (Equation 9).

$$MRP = \frac{(\text{PAIR} - \text{PAAR}) * 100}{PAIR} \quad (9)$$

Where, PAIR = Polygon area bordering the initial misaligned road vector layer and real road and PAAR = Polygon area bordering the aligned road vector layer and real road.

Therefore, higher MRP value indicates higher accuracy of the alignment approach while lower MRP value indicates lower accuracy of the alignment approach as the objective of the alignment approaches are to reduce the polygon area between the real road and the aligned road vector layer.

4. RESULTS

Considering the values obtained for the polygon area of test scenarios, the MRP was calculated. The average MRP values for the test scenarios 1, 2, 3, 4 is shown below in Table 1. Special road segments where the misaligned road segment has a lower value for the polygon area than the aligned road segments of all approaches were not considered for the MRP calculation.

Table 1. Average MRP for all test scenarios.

Scenario	Bézier Curves Order 2	Bézier Curves Order 3	Piecewise Linear Rubber Sheeting
1	71.24	68.64	69.13
2	76.58	80.70	78.48
3	52.70	71.08	63.70
4	72.83	64.11	67.88

Generally, the polygon area between the ground truth data and the aligned road segment was reduced for all alignment approaches for majority of road segments in each test scenario. Thus, resulting in positive MRP values. However, several road segments stood out as exceptions. For road segment 1.5 (Segment 5 in test scenario 1) and 2.1, the piecewise linear rubber sheeting approach with delaunay triangulation could not be applied as the road segment is situated out of the triangles generated from the provided control points whereas the other approaches have reduced the polygon area. This can be avoided by providing an additional control point such that the road segment is situated within a triangle. The road segment 2.6 was a special situation where the misaligned road segment has a lower value for the polygon area than the aligned road segments of all approaches. However, it can be observed that the misaligned vector layer has cut off the original road segment in multiple places in this situation. Road segments 3.1, 3.7 and 4.2 suffered from the same problem.

We observed that segment 3.3 has a significantly lower MRP for alignment considering bézier curves of order two approach. However, the MRP has significantly increased when the alignment considering bézier curves of order three approach is considered. This is due to the shape of the road segment which cannot be modeled by an order two bézier curve. It is noted that the bézier approach deforms the shape of the existing vector layer when the shape of a road segment cannot be modeled considering an order two or order three curve. However, the positional misalignment between the real road and the vector layer is reduced. When compared with the existing piecewise linear rubber sheeting approach using MRP values, order two and three bézier curve approaches respectively perform better for 11 and 13 road segments in total.

Table 2. Average MRP for all alignment approaches.

Approach	Average MRP
Piecewise Linear Rubber Sheeting	68.6
Bezier Approach (Order 2)	67.4
Bezier Approach (Order 3)	70.7
Bezier Approach (Combined)	75.8

Average MRP values for each approach is shown in Table 2. A combined average is derived for the bézier approach by obtaining the maximum MRP for each road segment out of order two and order three bézier approaches and then calculating the average of resulting figures. The combined bézier approach has a higher average MRP value of 75.8 than all other approaches. Whereas the bézier approach considering order two curves has the lowest average MRP value of 67.4. This showcases the ability of the proposed approaches to be used over the existing piecewise linear rubber sheeting approach in required situations to obtain a desired conflation between the road vector layer and the road layer in the satellite imagery.

5. CONCLUSION AND FUTURE WORK

This research contributes to the domain of vector to imagery conflation by introducing a new conflation algorithm. The proposed approach is capable of handling misalignments of a road segment in the vector format with a minimum number of two control points while the existing piecewise linear rubber sheeting approach with delaunay triangulation requires a minimum number of three control points to generate triangles. Also, the results showcase the proposed approach have a higher misalignment reduction percentage than the piecewise linear rubber-sheeting approach. In comparison with the existing snakes related approach for alignment, the proposed approach does not require extracting road features (road edges) for the alignment. The research also contributed to the domain of GIS by introducing a new evaluation model which can be utilized to measure the success of conflation approaches as well as road extraction from imagery approaches. The proposed alignment approach in this research as well as the existing alignment approaches as Snakes algorithm and Piecewise Linear Rubber Sheeting approach does not perform well when the existing road vector layer shape and the actual road layer shape is inconsistent from each other. Therefore, an approach to align the vector layer by providing additional intermediate control points in such situations can be explored. Further, in this research the road is considered as a quadratic or a cubic bézier curve to carry out the alignment process. Evaluation results showcased that when order two and order three curves were combined the misalignment reduction percentage increased. Thus, this approach can be generalized for any type of road by proposing a generalized curvature equation for all road layers or determining the number of control points required to generate the curve of the road when the points of the vector layer is given. Therefore, providing the capability of easily transforming any type of road layer without deforming the shape of the existing vector layer.

6. REFERENCES

[1] Andrášik, R., and Bíl, M. 2016. Efficient road geometry identification from digital vector data. *J. Geogr. Syst.* 18, 3 (01 Jul 2016), 249–264.

[2] Cao, C., and Sun, Y. Automatic road centerline extraction from imagery using road gps data. *Remote Sensing* 6, 9 (2014), 9014–9033.

[3] Chen, C.-C., Knoblock, C. A., and Shahabi, C. Automatically conflating road vector data with orthoimagery. *GeoInformatica* 10, 4 (2006), 495–530.

[4] Chen, C.-C., Thakkar, S., Knoblock, C., and Shahabi, C. Automatically annotating and integrating spatial datasets. In *International symposium on spatial and temporal databases* (2003), Springer, pp. 469–488.

[5] Fischler, M. A., and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

[6] Hwang, J.-R., Oh, J.-H., and Li, K.-J. Query transformation method by delaunay triangulation for multi-source distributed spatial database systems. In *Proceedings of the 9th ACM international symposium on Advances in geographic information systems* (2001), ACM, pp. 41–46.

[7] Kamermans, M. 2018. A Primer on Bézier Curves. https://pomax. github.io/bezierinfo/

[8] Maboudi, M., Amini, J., Hahn, M., and Saati, M. 2016. Road Network Extraction from VHR Satellite Images Using Context Aware Object Feature Integration and Tensor Voting. *Remote Sensing* 8, 8 (2016).

[9] Máttyus, G., Wang, S., Fidler, S., and Urtasun, R. 2015. Enhancing Road Maps by Parsing Aerial Images Around the World. In *IEEE I. Conf. Comp. Vis. (ICCV)*. 1689–1697.

[10] Ruiz, J. J., Ariza, F. J., Urena, M. A., and Blázquez, E. B. Digital map conflation: a review of the process and a proposal for classification. *Int. J. Geogr. Inf. Sci.* 25, 9 (2011), 1439–1466.

[11] Sederberg, T. W. Computer aided geometric design.

[12] Song, W., Keller, J. M., Haithcoat, T. L., and Davis, C. H. Automated geospatial conflation of vector road maps to high resolution imagery. *IEEE T. Image process.* 18, 2 (2009), 388–400.

[13] Song, W., Keller, J. M., Haithcoat, T. L., and Davis, C. H. Relaxation-based point feature matching for vector map conflation. *T. GIS* 15, 1 (2011), 43–60.

[14] Song, W., Keller, J. M., Haithcoat, T. L., Davis, C. H., and Hinsen, J. B. An automated approach for the conflation of vector parcel map with imagery. *Photogramm. Eng. Rem. S.* 79, 6 (2013), 535–543.

[15] Walter, V., and Fritsch, D. Matching spatial data sets: a statistical approach. *Int. J. Geogr. Inf. Sci.* 13, 5 (1999), 445–473.

[16] White Jr, M. S., and Griffin, P. Piecewise linear rubber-sheet map transformation. *The American Cartographer* 12, 2 (1985), 123–131.

Single Depth Map Super-resolution with Local Self-similarity

Xiaochuan Wang
State Key Laboratory of VRTS
Beihang University
Beijing, China
wangxc@buaa.edu.cn

Kai Wang
State Key Laboratory of VRTS
Beihang University
Beijing, China
wangkai_like@buaa.edu.cn

Xiaohui Liang
State Key Laboratory of VRTS
Beihang University
Beijing, China
Liang_xiaohui@buaa.edu.cn

ABSTRACT
Consumer depth sensors such as time-of-flight camera or Kinect have gained significant popularity in recently. However, the captured depth maps suffer from limited spatial resolution and a variety of noise, making such depth maps difficult to be directly applied in related applications. In this paper, we present a novel single depth map super-resolution method, aiming to reconstruct high-resolution depth map from its associated low-resolution depth map without any auxiliary information. Particularly, we exploit the depth local self-similarity to assist in constructing patch pairs in terms of high-resolution and low-resolution depth edge patches, and then deduce a high-resolution depth edge map via Markov model. Finally, we implement a joint bilateral filter to reconstruct the high-resolution depth map. Experimental results show that our method overcomes existing methods on the benchmark database as well as Kinect captured depth maps.

CCS Concepts
• **Computing methodologies** → **Image processing; Image-based rendering**; • **Computing methodologies** →**Multimedia streaming**

Keywords
Depth map; Super-resolution; Depth local self-similarity; Markov model

1. INTRODUCTION
With the development of consumer depth sensors such as Time-of-Flight (ToF) and structure light cameras [5], depth map becomes popular and thus is widely used in free-viewpoint video [16], 3D reconstruction [8], person re-identification [11] and other computer vision tasks. Due to the limitation of depth sensors, the captured depth map suffers from low spatial resolution and noise, making it incapable to support above applications directly. Therefore, depth super-resolution (SR) becomes necessary.

As is known, depth SR is an ill-posed problem. Most existing depth SR methods follow traditional image SR methods but introduce auxiliary information to alleviate reconstruction distortions.

ICVIP 2018, December 29-31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301515

Schuon et al. [15] proposed to utilize multiple low-resolution (LR) depth images from adjacent viewpoints to reconstruct a HR one. Cui et al. [2] followed it by allowing larger viewpoint baseline. Those methods require multi-view LR depth maps, which are notalways available in practical. Despite multiple LR depth maps, HR intensity image can also be employed for depth SR, which is first introduced by Diebel et al. [3]. It assumes that discontinuities in LR depth map are related to the edges in its associated HR intensity image. Yang et al. [18] applied a cross bilateral filter to refine the reconstructed depth map iteratively, where the intensity image is used to build a cost volume. Park et al. [13] extended [18] by using a nonlocal mean filter to preserve thin structure. Choi et al. [1] proposed a region segmentation based method to address texture transfer and depth bleeding artifacts. As addressed before, HR intensity image benefits depth SR, but induces new distortions, like texture transfer or texture copy.

Recently, single depth SR method become a spotlight. Aodha et al. [10] first proposed a patch based Markov model for single depth reconstruction. Xie et al. [17] extended the framework by transforming depth reconstruction from the image texture-edge prediction problem. In Ferstl et al. [4] a sparse coding is utilized to predict depth discontinuities in HR domain with dictionaries trained on the synthetic depth maps. Riegler et al. [14] propose a deep learning method, namely ATGV-Net, which comprises data-driven methods and energy minimization methods. Those methods, however, rely on large amount of HR-LR depth map pairs, which is difficult to construct in practical. Other than constructing large external training depth map pairs, Hornacek et al. [7] extend [10] but used patches only from the input depth maps. Li et al. [9] propose a similar framework by adding geometric constraints. These methods alleviate heavy training data, however, suffer from inaccurate patch mapping due to insufficient training sets.

Haven observed that depth map follows a local self-similarity prior, we extend exiting patch-based single depth map SR framework. Our contributions include:

- Derived from image self-similarity prior, we validate that depth map follow a similar prior that small patches are similar to themselves with different scaling factors.
- By using the depth local self-similarity prior, we extend exiting single depth reconstruction framework, where HR-LR patch mapping is refined.
- We validate our method on MiddleBury Stereo dataset and Kinect captured depth maps, demonstrating that our method outperforms related methods.

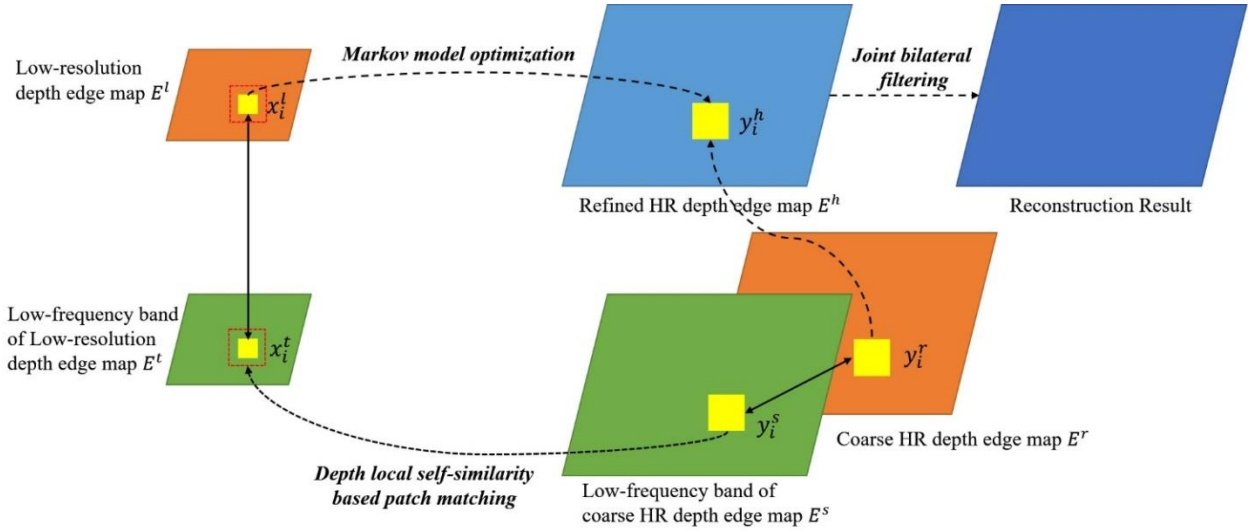


Figure 1. Overview of proposed method. $\mathbf{E^l}$, $\mathbf{E^t}$, $\mathbf{E^s}$, $\mathbf{E^r}$ are preprocessed HR-LR depth edge map pairs (Section 2.2), yellow boxes and red dotted boxes indicates patch matching based on depth local self-similarity. The Markov model optimization is applied to generate the refined HR depth edge map, followed a joint bilateral filtering.

2. APPROACH

2.1 Overview

Our work focuses on upscaling single depth map by utilizing its low-frequency content. Specially, our approach can be divided into three processes. Given a LR depth map, we first construct two HR-LR depth map pairs (Section 2.2), and then employ depth local self-similarity to find candidate patches (Section 2.3). Finally, we reconstruct the HR depth map from the edge map via optimization of Markov model (Section 2.4).

2.2 Depth Map Preprocessing

Provided an LR depth map D^l, we first obtain its low-frequency band D^t through Gaussian filtering. We then extract edge maps E^l and E^t toward these two maps using Canny operator, respectively. The raw LR depth map is simply resized to the target resolution, where the coarsely reconstructed depth map is depicted as D^r. We also extract its edge map E^r. Meanwhile, we apply a Shock filter to D^r and extract the edge map E^s. As addressed in [12], Shock filter preserves low-frequency content. By doing so, two HR-LR edge map pairs $\mathrm{E}^l - E^r$ and $\mathrm{E}^t - E^s$ are constructed. The first pair presents the raw information, where the last pair preserves low-frequency information. The whole pre-processing is illustrated in Fig. 1.

2.3 Depth Local Self-similarity Based Patch Matching

Instead of deducing HR depth map directly as previous methods, we propose constructing the HR edge map firstly. Particularly, for each pixel p_i in E^r, we extract patches of size $\mathrm{w} \times \mathrm{w}$ in E^r and E^s centered at p^i respectively, denoted as y^r and y^s. The patch pair $\mathrm{Y} = \{\mathrm{y}^\mathrm{r}, \mathrm{y}^\mathrm{s}\}$ thereby form a HR exemplar. We then find the corresponding LR exemplar in E^l and E^t based on depth local self-similarity prior, forming a LR exemplar $\mathrm{X} = \{\mathrm{x}^\mathrm{l}, \mathrm{x}^\mathrm{t}\}$.

Previous works have validated that natural image obeys a local self-similarity prior. i.e., image patches are similar to their surroundings within the same image. Freeman et al. [6] firstly introduced image self-similarity into image super-resolution, where the relevant patch exemplar is searched in neighborhoods around the relative coordinates in the input image. Depth map is different from natural image, however, follows the similar prior that depth singular features, e.g., discontinuities, are invariable when the depth patch is scaled with small factors. We call this property depth local self-similarity prior. It implies that relevant HR depth patches can be retrieved at a very restricted set of patches from LR versions at localized regions.

With the observed depth local self-similarity prior, we propose reconstructing HR edge map by itself instead of using external HR patches. Compared with existing exemplar or CNN based SR methods, the patches found in the input depth map itself are more relevant and natural, thereby ensuring comparable or even superior reconstruction results. However, no external dataset or HR intensity image are required.

Figure. 2 shows how well the depth local self-similarity holds at various scaling factors. Supposing a list of query patches, we retrieve their associated patches with different scaling factors. The l_1 norm error between query and retrieved patches is calculated. As can be seen, the error grows slowly when the scaling factor is small. However, the error becomes larger with the growing of scaling factor. This can be explained as follows. Larger scaling factors induce smoothness when computing the pixel errors. The singular features, especially edges are no longer scale-invariable then. To alleviate mismatching, we adopt an iterative strategy in our SR process. For each iteration, the patch pair with a small scaling factor is matched.

To validate the depth local self-similarity in patch matching, we quantify the matching error between the query depth patch and the retrieved depth patches centering around the original location the query patch. Figure. 3 shows the average matching errors among the MiddleBury Stereo dataset, where the query patch is 7×7 and the retrieve window is 9×9. The "yellow" regions denote large matching errors and the "blue" regions present small matching errors. As can be seen, the closer to the original location, the lower the matching error induced. The lower matching error

occurs at the very center on average. Therefore, we can find a well matched patch for a query depth patch localized around its original location, thereby reducing retrieval time in comparison with exhaust search.

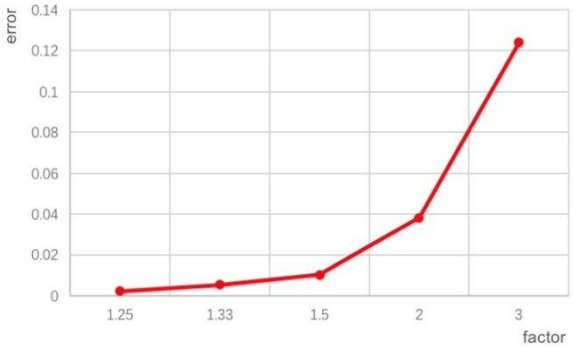


Figure 2. Errors between query patches and associated patches with different scaling factors. The patches are chosen from MiddleBury Stereo dataset.

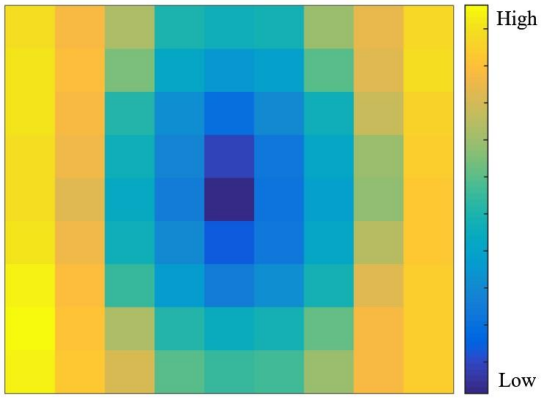


Figure 3. Errors between query patches and associated patches with different scaling factors.

2.4 Depth Reconstruction Via Markov Model Optimization

Suppose the matched HR-LR pair exemplar $\{Y, X\}$, we refine the coarse HR depth edge map by minimizing a discrete Markov model energy function. Particularly, each LR patches x_i forms nodes and their associated HR patches y_i forms hidden labels. The total energy function is formulated as follows.

$$E(y) = E_1 + \omega_1 E_2 + \omega_2 E_3 \quad (1)$$

where $E(y)$ indicates the label set of the reconstructed HR depth edge map, E_1 and E_2 are data terms, and E_3 is the smoothness term.

The first data term E_1 encodes the likelihood of the similarity between query depth patch x^l in LR depth edge map and the candidate patches y^r in coarse HR depth edge map in terms of Euclidean distance of their corresponding distance transforms:

$$E_1 = \sum_{i=1}^{N} \left\| d(x_i^l) - d(y_i^r) \right\|^2 \quad (2)$$

Where $d(\cdot)$ indicates the distance transform, which provides accurate similarity measurement in terms of binary patterns. The second data term E_2 ensures that the HR edge patch candidates have consistent similarity measurement in terms of the low-frequency edge pattern:

$$E_2 = \sum_{i=1}^{N} \| x_i^s - y_i^s \|^2 \quad (3)$$

The smoothness term E_3 enforces coherence of overlapping regions between neighboring patch candidates, where $O_{\{ij\}}$ is an overlapping operator that extracts overlapped region between two adjacent patches y_i^s and y_j^s:

$$E_3 = \sum_{x_i^l \cap x_j^l \neq \emptyset}^{N} \left\| O_{ij}(d(y_i^s)) - O_{ij}\left(d(y_j^s)\right) \right\|^2 \quad (4)$$

Particularly, for each $Y_i = \{y_i^r, y_i^s\}$, we find its N candidate patches in $X_i = \{x_i^l, x_i^s\}$ by using our depth local self-similarity prior, and then use belief propagation to minimize Eq.(1). The discrete label of each Y_i in terms of an edge patch in X_i can thus be inferred. Finally, y_i^s is put together by averaging pixel values in the overlapped regions, followed by applying threshold to the binary edge map.

With the refined HR depth edge map E^h, we reconstruct the HR depth map D^h through joint bilateral filtering.

3. EXPERIMENT RESULTS

In this section, we compare our proposed method with state-of-the-art depth image SR methods in both quantitative and visual ways. Particularly, we test the performance on the benchmark MiddleBury Stereo dataset as well as captured depth maps using Microsoft Kinect v2. In order to maintain the depth local self-similarity prior, we restrict the scaling factor to 2. As addressed before, the HR depth map is iteratively reconstructed. Suppose the target resolution is 2^n times the raw resolution, a total of n iterations are implemented. For the Markov model optimization, we set $\omega_1 = 8$ and $\omega_2 = 1$ in Eq.(1).

3.1 Quantitative Results

To evaluate the performance of our method, we choose 5 methods as comparisons, including Nearest Neighbor interpolation (NN), Yang et al. [18], Aodha et al. [10], Xie et al. [17], and ATGV-Net [14]. Among these methods, NN is traditional image interpolation, Yang et al. [18] is guided by HR intensity image, Aodna et al. [10] and Xie et al. [17] are typical single depth map SR methods. ATGV-Net [14] is CNN-based. In order to validate the effect of our proposed depth local self-similarity, we apply our framework but substitute the depth local self-similarity with image self-similarity prior [6], denoted as Freeman+.

Two indicators, including RMSE and SSIM are used to evaluate the SR performance. The former one represents average pixel errors between the reconstructed HR depth map and the ground truth. The latter one indicates the structure distortions of reconstructed depth map toward human perception. Lower RMSE and Higher SSIM values represent better reconstruction quality. Particularly, we choose the HR depth maps from MiddleBury Stereo dataset as the ground truth, and then down-sample them to obtain LR depth maps. The performance of our proposed method is shown in Table 1.

We can see from Table 1 that our method outperforms traditional depth SR methods, especially those utilize auxiliary information. Particularly, our method achieves competitive performance with ATGV-Net, while the former one needs heavy training samples and complex variation model optimization. Note that Freeman+ achieves modest performance compared with ours, mainly due to the different similarity assumptions.

3.2 Visual Results

The visual quality of reconstructed HR depth maps of *Teddy* and *Cones* with scaling factor of 4 is demonstrated in Fig. 4. As can be seen, Aodha et al. [10] and Xie et al. [17] induce artifacts especially around depth discontinuities. This is mainly due to the strategy that utilize external patch samples directly. In contrast, our method preserves visual structures by leveraging the depth local self-similarity. Figure. 5 shows another set of results with cropped zooming regions of *Tsukuba* and *Venus* with scaling factor of 4. We can see from Fig. 5 that the results of Aodha et al. [10] are over-sharp, inducing noticeable artifacts, while the results of Xie et al. [17] are and jaggy. Our results alleviate jaggy by refining the HR depth map with Markov model optimization.

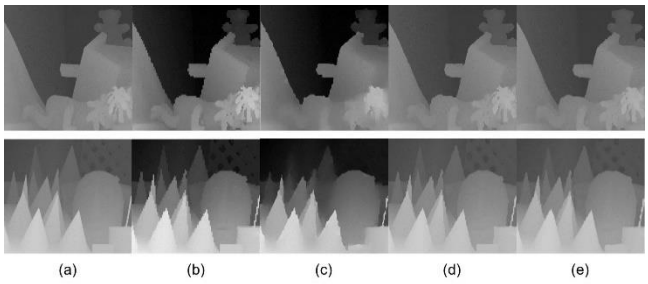


Figure 4. Visual quality of reconstructed depth map (×4) with different methods. The upper row is *Teddy*, and the bottom row is Cones from MiddleBury Stereo dataset. (a) is the ground truth, (b)-(d) are reconstructed results with nearest neighbor (NN), Aodha [10], and Xie [17]. (e) is our results.

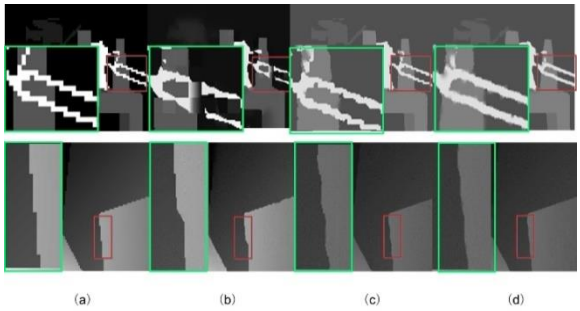


Figure 5. Visual quality of reconstructed depth map. The upper row is *Tsukuba* and the bottom row is *Venus*. Pixels in the red boxes are zoomed in and shown in the greed boxes.

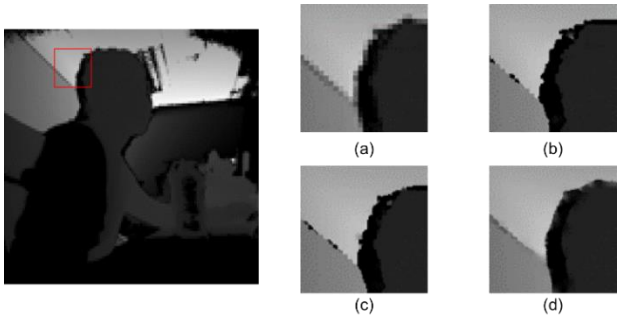


Figure 6. Visual quality of reconstructed depth map from noisy LR depth map captured by Kinect v2(×4). The left is the raw noisy LR depth map. (a-d) are the reconstructed results in terms of the red box to the left image. Particularly, (a)-(c) are reconstructed results by NN, Aodha [10], and Xie [17]. (d) is our results.

We also show the visual quality of reconstructed HR depth maps from Kinect captured noisy LR depth maps, as illustrated in Fig. 6. It can be seen that our method alleviates noise in the reconstructed depth map, while Aodha et al. [10] and Xie et al. [17] produce irregular edges and artifacts. This is because the patch matching with external dataset is sensitive to noise, thereby inducing inaccurate patch pair mapping. Our method, however, ensures that the found patch candidates are always in the raw depth map. Moreover, the Markov model optimization enhances the reconstruction from HR edge map to the final HR depth map.

4. CONCLUSION

In this paper, we have proposed a novel method for single depth image super-resolution based on depth local self-similarity, reconstructing HR depth map from the LR depth map itself without any auxiliary information. Specially, we exploit depth local self-similarity, with which to establish HR-LR depth patch pairs.

Additionally, we employ the low-frequency information in the raw depth map, constructing two HR-LR depth edge pairs.

Table 1. RMSE and SSIM comparisons on the MiddleBury Stereo dataset with scaling factor (×4)

	RMSE				SSIM			
	Cones	Venus	Teddy	Tsukuba	Cones	Venus	Teddy	Tsukuba
NN	1.513	0.424	1.586	0.933	0.886	0.954	0.895	0.833
Yang et al.	2.192	1.174	1.862	0.942	0.868	0.924	0.873	0.777
Aodha et al.	1.496	0.389	1.506	0.934	0.922	0.961	0.902	0.839
Xie et al.	1.16	0.314	1.025	0.742	0.915	0.97	0.918	0.843
AGTV-Net	**1.002**	**0.199**	**0.815**	**0.717**	**0.93**	**0.977**	**0.928**	**0.871**
Freeman+	1.411	0.397	1.532	0.976	0.924	0.959	0.907	0.839
Ours	**1.052**	**0.248**	**0.888**	**0.723**	**0.926**	**0.976**	**0.927**	**0.853**

We then find the matched depth edge patches with depth local self-similarity prior. After obtaining the coarse HR depth edge map, we further refine it with a Markov model optimization. Finally, we get the HR depth map with joint bilateral filtering. Compared with existing methods, our method does not need guidance of HR intensity image or extern patch dataset, while reducing blurring and jaggy artifacts in the reconstructed depth maps.

5. ACKNOWLEDGMENTS

This work is sponsored by the National Key R\&D Program of China (No.2017YFB1002702) and the National Nature Science Foundation of China (No.61572058).

6. REFERENCES

[1] O. Choi and S.-W. Jung. A consensus-driven approach for structure and texture aware depth map upsampling. IEEE Transactions on Image Processing, 23(8):3321–3335, 2014.

[2] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 1173–1180. IEEE, 2010.

[3] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In Advances in neural information processing systems, pages 291–298, 2006.

[4] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof. Image guided depth upsampling using anisotropic total generalized variation. In Proceedings of the IEEE International Conference on Computer Vision, pages 993–1000, 2013.

[5] S. Foix, G. Alenya, and C. Torras. Lock-in time-of-flight (tof) cameras: A survey. IEEE Sensors Journal, 11(9):1917–1926, 2011.

[6] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. IEEE Computer graphics and Applications, 22(2):56–65, 2002.

[7] M. Hornácek, C. Rhemann, M. Gelautz, and C. Rother. Depth super resolution by rigid body self-similarity in 3d. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1123–1130, 2013.

[8] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In Proceedings of the 24th annual ACM symposium on User interface software and technology, pages 559–568. ACM, 2011.

[9] J. Li, Z. Lu, G. Zeng, R. Gan, and H. Zha. Similarity-aware patchwork assembly for depth image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3374–3381, 2014.

[10] O. Mac Aodha, N. D. Campbell, A. Nair, and G. J. Brostow. Patch based synthesis for single depth image super-resolution. In European conference on computer vision, pages 71–84. Springer, 2012.

[11] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. Van Gool. One-shot person re-identification with a consumer depth camera. In Person Re-Identification, pages 161–181. Springer, 2014.

[12] S. Osher and L. I. Rudin. Feature-oriented image enhancement using shock filters. SIAM Journal on numerical analysis, 27(4):919–940, 1990.

[13] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 1623–1630. IEEE, 2011.

[14] G. Riegler, M. Rüther, and H. Bischof. Atgv-net: Accurate depth super-resolution. In European Conference on Computer Vision, pages 268–284. Springer, 2016.

[15] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 343–350. IEEE, 2009.

[16] A. Smolic. 3d video and free viewpoint video from capture to display. Pattern recognition, 44(9):1958–1968, 2011.

[17] J. Xie, R. S. Feris, and M.-T. Sun. Edge-guided single depth image super resolution. IEEE Transactions on Image Processing, 25(1):428–438, 2016.

[18] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007

Object Recognition Using Deep Neural Network with Distinctive Features

Hyun Chul Song
Department of Computer Science and Engineering
Chung-Ang University
Seoul, Korea
hcsong@vim.cau.ac.kr

Farhan Akram
Imaging Informatics Division,
Bioinformatics Institute,
Singapore.
farhana@bii.astar.edu.sg

Kwang Nam Choi*
Department of Computer Science and Engineering
Chung-Ang University
Seoul, Korea
knchoi@vim.cau.ac.kr

ABSTRACT
In this paper, a new object recognition method using statistically weighting Multi-Layer Perceptron (MLP) is proposed. It uses visual distinctive features, which are computed using Bag of Visual Words (BoVW) framework. The proposed method has the following three main steps. At first it represents the images into their respective co-occurrence matrices, which are vectorized using BoVW and gives distinctive features. Then it computes weights from the histograms of visual words for each class. Finally, the statistically weighting distinctive features are applied to the testing image set to find the object class. In the proposed method, we improved MLP by introducing the weighted visual words, which are extracted by sampling the patches from the current image. From the Caltech 256 dataset, four classes namely pedestrians, cars, motorbikes and airplanes are used for the classification accuracy comparison between the MLP based artificial neural network (ANN) and the proposed method. The experimental results show that our method outperforms traditional MLP yielding an average classification accuracy of 89.60%, which is approximately 6.3% more than the compared MLP.

CCS Concepts
•Computing methodologies → Machine learning → Machine learning approaches → Neural networks

Keywords
Object recognition; bag of visual words; multi-layer perceptron; scale-invariant feature transform; weighting scheme

1. INTRODUCTION
Object or image recognition is a well-known problem in computer vision. It is process of taking an image and recognizing the features or structures of the image. After that classifying or grouping the input image or object into their respective group, for example, classifying orange and apple into different groups after knowing their features or characteristics.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12… $15.00
https://doi.org/10.1145/3301506.3301518

Recognizing pedestrian and transport objects is an essential and significant task in intelligent systems such as, video surveillance. It can be used as an extension to the automotive applications such as, autonomous vehicles including drones to improve safety systems. Nowadays, an active research is going to extract objects from an image and then organize them into different categories using machine learning [9, 12, 19, 7, 2]. Numerous models have been devised for object recognition and classification based on different methodologies. Bag of Visual Words (BoVW) is one of those models used for object recognition and classification.

BoVW generates a bag of visually distinctive features from the image by defining that image is a document and a visual word is a word obtained by information retrieval in the text domain. BoVW model simplifies data to analyze the representation of a co-occurrence matrix, in which columns represent indices and rows contain word count information of a document. Although it does not consider any spatial relation between the feature points, it still yields successful categorization in the textual domain. In the visual domain, topics are discovered by training data using a large set of unsupervised documents. Accordingly, the model used in text domain exploits the category of objects in the visual domain. In order to apply the model in the visual domain, a document is created that corresponds to an image and a visual word that quantizes the local feature descriptors of the image. This approach builds a vocabulary of visual keywords by clustering extracted feature vectors, which builds a histogram to compute the number of keyword occurrences in the image and then feed it to the classifier.

In this work, first a BoVW framework is used to generate visual distinctive features. Then an improved feature descriptor is proposed using artificial neural network (ANN) [11, 4], which is trained for BoVW features. The pro-posed feature descriptor back propagates the classification errors to the BoVW framework to improve the feature ex-traction function. The proposed object recognition method is tested by using a randomly initialized feature descriptor to determine the effect of initialization of existing feature descriptors on its performance. Finally, the original enhanced descriptor is tested with a different set of trained settings to see if it can be generalized for another recognition or classification framework. Fig. 1 shows a flow chat of the proposed recognition system. It shows that BoVW is a combination of Canny edge detector, Scale-Invariant Feature Transform(SIFT) based feature descriptor and clustering algorithm that formulates the visual words dictionary. In this paper, we use MLP to design a feature descriptor for BoVW distinctive features. Four classes i.e., pedestrian, cars, motorbikes and airplanes are used from Caltech 256 image database [6, 8] to generate and compare the classification results with traditional MLP.

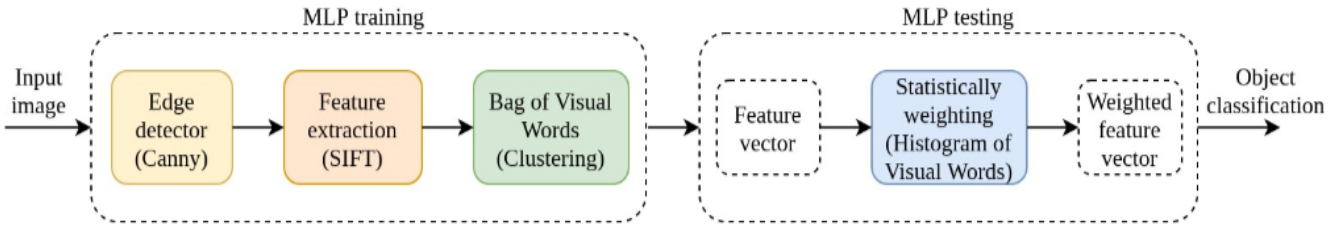


Figure 1: The proposed object classification framework

Figure 2: The Canny Edge Detector discovers the feature points to decrypt.

2. PREVIOUS WORK

Visual words are described by the points, which are computed by the various feature point detectors such as, image feature descriptors. In a visual words algorithm, a number of feature vectors obtained from a collection of images are clustered by using clustering algorithm and vector quantization. This information is later used by the bag of the visual words model; therefore, it can be said that visual words lays the foundations of the bag of visual words model. In this paper, a Canny edge detector is used as a feature point detector and SIFT descriptor is used as an image feature descriptor. In turn, the K-means clustering algorithm is used to formulate a visual wordbook based on the training data. Canny Edge Detector [10] is considered a power edge detection tool, which is widely used nowadays. It computes the largest count number of pixels in an edge by minimizing the number of edges with respect to reference pixels. Canny edges that show significantly different values among the input images are used to avoid the bias produced by the irrelevant edges. Fig.2 shows the feature point discovered by the Canny edge detector. SIFT descriptor is used for feature point description. It is used as a fundamental method to describe an object in various visual areas such as, object recognition and image reconstruction. It is robust against the changes in the view point, size, translation, rotation and illumination of an object. In particular, SIFT is quite effective in affine transform and 3D description. It defines the largest values of the difference of a Gaussian in a pyramid scale space as a point of interest. It can also be used as a descriptor with a histogram of direction factors, which shows reliable matching of the object or foreground using different viewpoints. It has a low time complexity and can be used for visually discriminative features in the object matching by selecting the area of interest defined by its current state [17].

BoVW framework is inspired by the bag of words method, which is often used in the text classification [14, 16]. In BoVW, word frequency information is collected by using text classification and stored in the histogram Numerous classification methods are used to determine the semantic context of the text using histogram oriented schemes [22, 1, 3, 5, 15, 20]. BoVW framework is used to cluster the features extracted from the training image to obtain the visual keywords. The extracted characteristics of a given testing image match the visual keywords. Later, the number of matches per cluster is stored in the histogram, which is used as an input to the classifier. This cluster matching is called hard bag function because it uses hard allocation method [22]. In turn, a different approach is also introduced that uses a soft allocation to construct a histogram. It weights multiple clusters close to their capabilities. The weights are defined by ranking the nearest neighbors in terms of distance from the central point [13] or by using the distance itself [21, 23].

3. BACK-PROPAGATION AND OTHER DIFFERENCIATION ALGORITHMS

3.1 ACTIVATION FUNCTIONS

In artificial neural networks, the activation function of a node defines the output of that node given an input or set of inputs. The Rectified Linear Unit (ReLU) is the most commonly used activation function in deep learning models. The function returns 0 if it receives any negative input, but for any positive value x it returns that value back. In this paper, ReLU activation function is used for all experiments.

3.2 BACK-PROPAGATION

Fully connected MLP is applied in a forward-propagation fashion to yield acceptable classification performance in testing phase. The scale of MLP consists of the depth and numbers of neurons of each layer. The depth is a number of hidden layers and an output layer. Fig. 3 shows the network structure of proposed model. In this work, input feature vector size = 300 and number of output classes = 4 and hidden layers = 6.

In the training stage, the back-propagation is used to reduce the error between two energy differences by applying gradient descent method based on the least mean squares to update the weights of each layer. The cost function is a sum of error signals over *N* patterns, which is defined as

$$E_{av} = \frac{1}{N} \sum_{k=1}^{N} E(n). \quad (1)$$

By minimizing the cost function, the network learns the parameters through the example patterns. The sum of error signals on output layer with nth pattern is defined by

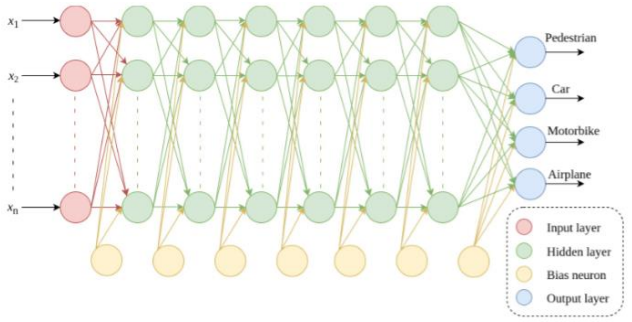


Figure 3: The architecture of MLP based NN used in the proposed classification method.

$$E(n) = \frac{1}{2}\sum_{k=1}^{m_L} e_k^2(n), \tag{2}$$

Where m_L is a number of neurons on output layer. The error signal $e_k(n)$ is a difference between the desired value $d_k(n)$ and the functional signal $y_k(n)$ in the output layer that is defined as

$$e_k(n) = d_k(n) - y_k(n). \tag{3}$$

The local field value goes through the activation function associated with the nonlinearity, which is given as

$$y_k(n) = \varphi(v_k(n)). \tag{4}$$

The weighted summation from previous layer of neuron K at iteration n is described as

$$v_k(n) = \sum_{j=1}^{m} w_{kj}(n) y_j(n), \tag{5}$$

Where w_{kj} is a weight from neuron j to neuron k and m is a number of the previous layer. The back-propagation method updates the weight $w(n)$ in a similar fashion as Least mean squares (LMS) algorithm, as defined below

$$w(n+1) = w(n) + \Delta w(n), \tag{6}$$

There are two cases in which weight $\Delta w(n)$ is minimized for both output and hidden layers. In the first case, minimization Δw_{kj} for the output layer k in iteration n is defined by the delta rule and the chain rule of calculus:

$$\begin{aligned}\Delta w_{kj}(n) &= -\eta\frac{\partial E(n)}{\partial w_{kj}(n)} \\ &= -\eta\frac{\partial E(n)}{\partial e_k(n)}\frac{\partial e_k(n)}{\partial y_k(n)}\frac{\partial y_k(n)}{\partial v_k(n)}\frac{\partial v_k(n)}{\partial w_{kj}(n)} \\ &= \eta e_k(n)\varphi\prime_k(v_k(n))y_j(n) \\ &= \eta\delta_k(n)y_j(n),\end{aligned} \tag{7}$$

Where η is the learning rate. The $\delta_k(n)$ is a local gradient applied to weight w_{kj} as defined by

$$\begin{aligned}\delta_k(n) &= -\frac{\partial E(n)}{\partial v_k(n)} \\ &= -\frac{\partial E(n)}{\partial e_k(n)}\frac{\partial e_k(n)}{\partial y_k(n)}\frac{\partial y_k(n)}{\partial v_k(n)} \\ &= e_k(n)\varphi\prime_k(v_k(n))\end{aligned} \tag{8}$$

In the second case, minimization δw_{ji} to update w_{ji} for the hidden layer j as defined by

$$\begin{aligned}\Delta w_{ji}(n) &= -\eta\frac{\partial E(n)}{\partial w_{ji}(n)} \\ &= -\eta\sum_{k=1}^{m_L} e_k\frac{\partial e_k(n)}{\partial y_k(n)}\frac{\partial y_k(n)}{\partial v_k(n)}\frac{\partial v_k(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial v_j(n)}\frac{\partial v_j(n)}{\partial w_{ji}(n)} \\ &= \eta\varphi\prime_j(v_j(n))y_i(n)\sum_{k=1}^{m_L} e_k(n)\varphi\prime_k(v_k(n))w_{ji}(n) \\ &= \eta\delta_j(n)y_i(n),\end{aligned} \tag{9}$$

The local gradient $\delta_j(n)$ is defined as

$$\delta_j(n) = \varphi_j(v_j(n))\sum_{k=1}^{m_L} e_k(n)\varphi\prime_k(v_k(n))w_{ji} \tag{10}$$

Let h be the previous layer and j the next layer to the hidden layer I, then minimization $\Delta w_{ih}(n)$ is back-propagated

recursively, which is defined as

$$\Delta w_{ih}(n) = -\eta\frac{\partial E(n)}{\partial w_{ih}(n)} = \eta\delta_i(n)y_h(n), \tag{11}$$

where the local gradient $\delta_i(n)$ is defined as

$$\delta_i(n) = \varphi\prime_i(v_i(n))\sum_{j=1}^{m_j}\sum_{k=1}^{m_L} e_k(n)\varphi\prime_k(v_k(n))w_{kj}\varphi\prime_j(v_j(n))w_{ji}(n) \tag{12}$$

The minimization $\Delta w(n)$ is computed by the product of local gradient and induced local field with learning rate during the back-propagation.

4. PROPOSED WEIGHTING FEATURE

Let N be the number of documents (i.e., images) and M major vocabularies in each document, which are defined as $D=d_1,d_2,...,d_N$ and $W=w_1,w_2,...,w_M$, respectively. By ignoring the sequential positional relationship in the document it can be arranged as a co-occurrence matrix of size $N \times M$, with entry $n(d_i,w_j)$ listing the number of words j in document i [22]. It can be represented as a vector space by being expressed as a bag of visual words and employed as the input feature vector of MLP. It shows feed forward and backward propagations in MLP structure with softmax. The proposed weights reflect the difference of distribution between one category of visual words to another. Visual words that appear dominantly in a specific category have higher values in the visual word histogram, while the common ones that appear in other categories have lower values. Histogram values can be used for layer weighting in order to increase the discriminative power of the algorithm. The weights from the difference of histogram are computed to multiply with the histogram of visual words. We obtained the weighted feature vectors which are multiplied by weights for each visual word.

$$Weights_{class_i} = \frac{(X_{class\ i} - E(X_{visual\ word\ j}))^2}{std(X_{visual\ word\ j})} \times r, \tag{13}$$

Table 1: Comparison between traditional MLP and the proposed method (improved MLP) using Caltech256 and NICTA datasets.

Class	MLP	Proposed Method
Pedestrians	84.0 %	98.0 %
Cars	69.0 %	85.1 %
Motorbikes	84.0 %	86.1 %
Airplanes	96.0 %	89.1 %
Average	83.3 %	89.6 %

Where r is a weight ratio and X visual word i stands for the i^{th} visual vocabulary over the classes.

$$\overline{X} = X \odot Max(Weights_{class_i}), \text{for } i = 0, ..., N. \quad (14)$$

The proposed method employs softmax and cross entropy for the output units. The cross entropy on output layer with n^{th} pattern is defined by

$$E(n) = \sum_{k=1}^{m_L} e_k(n), \quad (15)$$

Where m_L is a number of neurons in the output layer. The error signal $e_k(n)$ is a difference between the desired value $d_k(n)$ and functional signal $y_k(n)$ in the output layer, which is defined as

$$e_k(n) = -d_k(n) \log(y_k(n)). \quad (16)$$

The local field value acts as an input to the activation func-tion softmax in the output layer and ReLU in the hiddenlayer as given in Eq. 5. The softmax is defined as

$$y_k(n) = \frac{e^{v_k(n)}}{\sum_{c=1}^{m_L} e^{v_c(n)}}. \quad (17)$$

ReLU is the activation function of the hidden layer, which is defined as $max(0, v_k(n))$. Weight error minimization $\Delta w(n)$ is defined by the delta rule and the chain rule of calculus:

$$\begin{aligned} \frac{\partial E(n)}{\partial y_k(n)} &= \frac{\partial E(n)}{\partial e_k(n)} \frac{\partial e_k(n)}{\partial y_k(n)} = -\sum_{k=1}^{m_L} \frac{d_k(n)}{y_k(n)} \\ \frac{\partial y_l(n)}{\partial v_k(n)} &= \frac{\partial}{\partial v_k(n)} \frac{e^{v_l(n)}}{\sum_{c=1}^{m_L} e^{v_c(n)}} \end{aligned} \quad (18)$$

Let $f(x) = \frac{g(x)}{h(x)}$ the first order differentiation is $f\prime(x) = \frac{f\prime(x)g(x) - f(x)g\prime(x)}{(h(x))^2}$. Similiary, weight error minimization Δw_{kj} is defined

$$\begin{aligned} \Delta w_{kj}(n) &= -\eta \frac{\partial E(n)}{\partial w_{kj}(n)} = -\eta \frac{\partial E(n)}{\partial v_k(n)} \frac{\partial v_k(n)}{\partial w_{kj}(n)} \\ &= \eta (d_k(n) - y_k(n)) y_j(n) \\ &= \eta \delta_k(n) y_j(n), \end{aligned} \quad (19)$$

Where η is the learning rate. The $\delta_k(n)$ is a local gradient applied to weight w_{kj} as defined as

$$\delta_k(n) = -\frac{\partial E(n)}{\partial v_k(n)} = d_k(n) - y_k(n)) \quad (20)$$

The weight w_{ji} on the hidden layer j (second last layer) is defined as

$$\begin{aligned} \Delta w_{ij}(n) &= -\eta \frac{\partial E(n)}{\partial w_{ij}(n)} \\ &= \eta \varphi\prime_j(v_j(n)) \sum_{k=1}^{m_L} (d_k(n) - y_k(n)) w_{kj}(n) y_i(n) \\ &= \eta \delta_j(n) y_i(n), \end{aligned} \quad (21)$$

The local gradient $\delta_j(n)$ is defined as

$$\delta_j(n) = \varphi\prime_j(v_j(n)) \sum_{k=1}^{m_L} \delta_k w_{kj}(n) \quad (22)$$

The bias b_j is updated by Δb_j which is defined as

$$\Delta b_j(n) = -\eta \frac{\partial E(n)}{\partial b_j(n)} = \eta \delta_k(n). \quad (23)$$

Let h be the previous layer and j the next layer to the layer i, then weight error minimization $\Delta w_{ih}(n)$ is back-propagated recursively. By using weight error minimization from Eq. 12, local gradient $\delta_i(n)$ is defined as

$$\delta_i(n) = \varphi\prime_i(v_i(n)) \sum_{j=1}^{m_j} \sum_{k=1}^{m_L} (d_k(n) - y_k(n)) w_{kj} \varphi\prime_j(v_j(n)) w_{ji}(n) \quad (24)$$

The bias b_i is updated by Δb_i, which is defined as

$$\Delta b_i(n) = -\eta \frac{\partial E(n)}{\partial b_j(n)} = \eta \delta_i(n). \quad (25)$$

5. EXPERIMENTS AND RESULTS

In this paper, a system using proposed object recognition method for images is proposed to achieve improvements over standard MLP. In this work, the images used for experiments are: motorbikes, cars, airplanes and pedestrians, which are obtained from the Caltech 256 image database [8]. Each image class contains 200 images in which 50% images are used for training and 50% are used for the test evaluation. The sequence of images used for both training and testing are taken at random. NICTA pedestrian dataset is employed to train and then evaluate the dictionary [18]. The proposed method is tested and compared with the traditional MLP using different image classes. Results show that the proposed method improved the recognition rate when it is hard to differentiate between two classes (two different categories of images). In turn, the proposed method yields similar recognition results compared to the traditional MLP when there is a distinct visual difference between two categories of images The experiments are cross-evaluated for all categories. In each experiment, for recognition two different image categories are used at a time. The result shows that proposed method shows better recognition rates than standard MLP. It indicates that weighting the distinctive features enforces the object categorization by heavily weighting the dominant features. The proposed weighting method can also be easily adopted in other detection and classification method. Table. 1 shows a class recognition rate comparison between the proposed method and the traditional MLP. It shows that the proposed method yields about 6.3% higher accuracy than the traditional MLP for all classes.

6. CONCLUSIONS

In this paper, object recognition method using deep neural neural network is proposed by combining the bag of visual model with a

statistically weighting mechanism. It uses MLP based ANN as a weighted feature descriptor for object classification. It demonstrates the significant role of the weighting features based on the visual distinctiveness, where the weights are obtained by computing the dominant distributions of visual words over the classes. The experimental results using a public dataset indicate that the proposed method outperforms traditional MLP in classifying pedestrian and transportation objects. It yields 6.3% higher accuracy than the compared conventional MLP. It improves the recognition performance between similar categories with a low category recognition rates by assigning the weights to visual words.

7. ACKNOWLEDGMENTS

This work was supported by the Technology development Program(S2601546) funded by the Ministry of SMEs and Startups(MSS, Korea)

8. REFERENCES

[1] A. Abdullah, R. C. Veltkamp, and M. A. Wiering. Ensembles of novel visual keywords descriptors for image categorization. *In Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on*, pages 1206–1211. IEEE, 2010.

[2] Z. Al-Zaydi, B. Vuksanovic, and I. Habeeb. Image processing based ambient context-aware people detection and counting. *Int. J. Mach. Learn. Comput.(IJMLC)*, 8(3):268–273, 2018.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *In Workshop on statistical learning in computer vision, ECCV, volume 1*, pages 1–2. Prague, 2004.

[4] Y. R. Devi, S. Sarojini. A survey on machine learning and statistical techniques in bankruptcy prediction. *Int. J. Mach. Learn. Comput.(IJMLC)*, 8(2):268–273, 2018.

[5] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation: Generative models and pdf-kernels. 2005.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding*, 106(1):59–70, 2007.

[7] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. *In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages1816–1823. IEEE, 2005.

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object ategory dataset. *In California Institute of Technology*, 2007.

[9] H. Han, Q. Han, X. Li, and J. Gu. Hierarchical spatial pyramid max pooling based on sift features and sparse coding for image classification. *IET Computer Vision*, 7(2):144–150, 2013.

[10] C. Harris and M. Stephens. A combined corner and edge detector. *In Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988.

[11] M. Heidarysafa, K. Kowsari, D. E. Brown, K. J. Meimandi, and L. E. Barnes. An improvement of data classification using random multimodel deep learning(rmdl). *Int. J. Mach. Learn. Comput.(IJMLC)*, 8(4):268–273, 2018.

[12] Z. Ji. Decoupling sparse coding with fusion of fisher vectors and scalable svms for large-scale visual recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 450–457, 2013.

[13] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *In Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 494–501. ACM, 2007.

[14] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. *Technical report, DTIC Document*, 1996.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *In Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[16] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *In European conference on machine learning*, pages 4–15. Springer, 1998.

[17] D. G. Lowe. Local feature view clustering for 3d object recognition. *In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[18] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson. A new pedestrian dataset for supervised learning. *In Intelligent Vehicles Symposium, 2008 IEEE*, pages 373–378. IEEE, 2008.

[19] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. *In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3498–3505. IEEE, 2012.

[20] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. *In Computer Vision and Pattern Recognition, 2007. CVPR'07.IEEE Conference on*, pages 1–8. IEEE, 2007.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. *In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[22] J. Sivic, A. Zisserman, et al. Video google: A text retrieval approach to object matching in videos. *In iccv*, volume 2, pages 1470–1477, 2003.

[23] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE transactions on pattern analysis and machine intelligence*, 32(7):1271–1283, 2010.

Intelligent Information Systems and Image Processing: A Novel Pan-Sharpening Technique Based on Multiscale Decomposition

Ahmad Al Smadi
Xidian University, China
Ahmadsmadi16@yahoo.com

Ahed Abugabah
Zayed University, UAE
Ahed.Abugabah@zu.ac.ae

ABSTRACT
Pansharpening is a technique belong to the image fusion field, the main goal of pansharpening is to enhance the Spatial resolution of multi-spectral image while preserving the spectral resolution. In this paper a simple pan-sharpening technique using saliency detection based on a guided filter is presented. In this technique the guided filter is used to build saliency detection algorithm. The saliency can be identified by making a saliency map that can be an outstanding part of the salient information of the image. we evaluate our technique by using some real and degraded data-sets then compare our technique with some existing methods in both subjective and objective aspects. The experimental results show that the proposed strategy can accomplish magnificent execution in both subjective and objective aspects.

CCS Concepts
• Computing methodologies~Image processing

Keywords
Pansharpening; Saliency detection; Guided filtering.

1. INTRODUCTION
The objective of image fusion is to produce a composite image by coordinating the integral data from numerous source images of a similar scene. The need of image fusion when there are numerous pictures of a scene, each exhibiting distinctive sort of data. Remote sensing image fusion is a critical branch of image fusion. If the spatial resolution is higher, the spectral resolution must be bring down in the remote sensing imaging process as indicated by the breaking points on the signal to noise ratio [1], There are numerous remote sensing satellites, for example, IKONOS, GEO, GF, and QuickBird that furnishes a high spatial resolution yet with low spectral resolution panchromatic (PAN) image and a high spectral resolution yet with low spatial resolution multi-spectral (MS) image simultaneously, nonetheless, the panchromatic image covers a wide wavelength go while the multi-spectral band covers a smaller phantom range. There are likewise a few applications inside the field of remote sensing that advantage from pan-sharpened imagery, for example, change detection, classification [2], vegetation identification, and lithology analysis [3]. The way toward combining multispectral image with panchromatic image to expand the spatial determination of multispectral image is called pansharpening of multispectral image [4].

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12...$15.00
https://doi.org/10.1145/3301506.3301523

As of late, an extensive number of remote detecting image fusion techniques have been proposed. As indicated by the distinctive phases of the combination procedure, these techniques can be partitioned into three classifications: pixel level [5], feature level [6] and decision level [7]. Pixel-level based methods are most popular at present, therefore, this paper is focused on that because the pixel level fusion can transfer a large portion of the helpful substance from source image to the fused image [8]. The multi-scale picture combination plans, which is decomposed the source images into spatial primitives at various spatial scales, at that point coordinate these primitives to shape another multi-scale portrayal, lastly apply a converse multi-scale change to reproduce the fused image. Precedents of this methodology are for example the Laplacian pyramid [9], and the contrast pyramid [10], however, These strategies may produce artifacts in the fused image.

The as of late presented Guided Filter [11] is a computationally productive, edge-saving interpretation variation administrator in light of a nearby straight model which stays away from the downsides of two-sided separating and different past methodologies. At the point when the information image likewise fills in as the guided image, the guided filter carries on like the edge protecting bilateral filter. Thus, the guided filter can smoothly dispense with little points of interest while recuperating bigger scale edges at the point when connected in an iterative system. In view of guided filtering, Alexander et al. [12] proposed a multi-scale image fusion schema. Iterative guided filtering is utilized to decompose the source images into base and detail layers. The vast majority of the multi-scale decomposition fusion strategies require in excess of two decomposition levels to get palatable results. Furthermore, these techniques are computationally costly, require more memory and in addition calculation time. Pyramid and wavelet-based techniques may cause antiquities around edges. In terms of the multi-scale decomposition process, the fusion methods which are based on edge preserving decomposition need edge preserving decomposition filters. In view of that, these filters are complicated to execute and take time.

Recently, Durga et al. [13] proposed a new image fusion method based on saliency detection and two-scale image decomposition (SSD) to solve the aforementioned problems. They used several images to test their method, however, to extract the salience information from integral source images, the saliency extraction

algorithm is an effective way to do that, the saliency map extracted by taking the absolute value of the difference between the mean filter and the median filter. Note that, the window sizes for mean and median filters were 35 and 3, respectively. Moreover, based on salience map, they generated a new weighted map construction. Shruti et al. [14] proposed a modified multiresolution analysis based pan-sharpening technique that is based on saliency detection of an image (PSD). They applied the pan-sharpening technique for three bands which are red, green, and blue independently then made a concatenation of the three fusion results of these bands to get the final fusion result. Furthermore, the window size of mean and median filters in their method were 100 and 3, respectively. Based on the previous works we modifying the pan-sharpening technique based on the guided filter to construct the saliency map. In addition, the guided filter is added to our strategy to decrease the antiquities.

The contributions of the proposed technique consisting of the six following steps: 1) Two scale image decomposition; 2) Employing the guided filter to construct the salience maps; we chose the guided filter because it has a good perform to remove the noise or artifact while preserving the edge information; 3) Salience detection, visual saliency indicates to the prominent and notable parts of an image which attract more visual consideration than the other parts of image [15, 17]: 4) The weight maps construction, the MS image and the PAN image contain information for the same object, in order to obtain a good fusion result, you must fuse the noteworthy information from the images, so the weight maps are important to achieve this; 5) Base and detail layers fusion; 6) Final fusion result.

The rest of this paper is organized as follows: Section 2, The guided filter is introduced in details. In Section 3, the proposed method is presented. Section 4, the experimental results are introduced. Section 5, we conclude this paper.

2. GUIDED FILTER

The guided filter is an interpretation variation filter in light of a local linear model [11]. Guided image filtering consisting of the following: 1) Input image I; 2) Guidance image Y; 3) Output image O. The local linear transformation between the input image and guidance image produces the output image which has the structure of the guidance image, and involves the main information from the input image. The output image O can be estimated by the following equation:

$$O_i = a_k Y_i + b_k \forall i \in w_k \quad (1)$$

where O_i, Y_i are the i_{th} pixel value of the output and the guidance images, respectively, and w_k denotes a squared window of size (2r +1)×(2r +1). a_k and b_k represent the linear coefficients which are constants in the window w_k , by minimizing the squared difference between the output image O and input image I, the linear coefficients can be computed as follows:

$$E(a_k, b_k) = \sum_{i \in w_k} ((a_k Y_i + b_k - I_i)^2 + \zeta a_k^2) \quad (2)$$

where ζ represents a regularization parameter which is determined by the user. The solution of previous equation can be given by linear regression [16]:

$$a_k = \frac{(1/|w| \sum_{i \in wk} Y_i I_i - \mu_K \bar{I}_K)}{(\delta_K^2 + \zeta)} \quad (3)$$

$$b_K = \bar{I}_K - a_k \mu_K \quad (4)$$

where μ_K and δ_K^2 are the mean and variance of Y, respectively in w_k, $|w|$ represents the number of pixels in w_k, and $\bar{I}_K$ denotes the mean of I in w_k. The output image O_i can be computed according to the equation (5), hence, it will change only when it is being computed with different values of w_k , however, in order to overcome this problem, all the possible coefficient values of a_k and b_K should be averaged, then, the output image O_i will be estimated as follows:

$$O_i = \bar{a}_i I_i + \bar{b}_i \quad (5)$$

where $\bar{a}_i = (1/|w|) \sum_{K \in w_i} a_k$ and $\bar{b}_i = (1/|w|) \sum_{K \in w_i} b_K$.

3. PROPOSED METHOD

The proposed strategy needs six stages to perform: decomposing the source images, salience detection, the weight maps construction, base and detail layers fusion, and final fusion result. These disintegrated base and detail layers are combined utilizing diverse combination rules. The combined image is remade from the final base and detail layers. the flow chart of the proposed technique appears in Fig. 1.

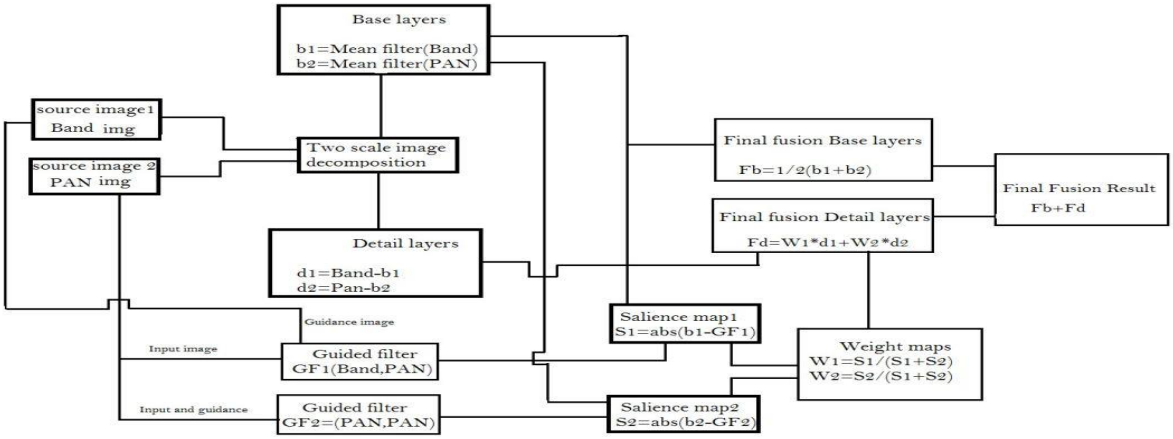


Figure 1. Flowchart of the proposed method.

3.1 Decomposing the Source Image

Two scale decomposition applied to the source images which are the MS and PAN images after one step which is up-sampling the MS image corresponding to the PAN image. In this step, we decompose the source images to base layers and detail layers. In term of the base layer contains the basic information or noticeable variance. The base layers are obtained by applying the mean filter on each band of the MS image and the PAN image. With regard to the detail layers which contain the tiny details, therefore, the detail layers can be obtained by subtracting the base layers from the source images. Mathematically, the following equations illustrate how can we generate the base and detail layers:

$$B_1 = M_{(x,y)} * I_{(x,y)} \tag{6}$$

$$B_2 = M_{(x,y)} * PAN_{(x,y)} \tag{7}$$

where $I_{(x,y)}, PAN_{(x,y)}$ represent the source images, $B_1 and B_2$ are the base layers of the first source image and PAN image, respectively. $M_{(x,y)}$ denotes to the mean filter of window size σ_M and $*$ represents the convolution. The detail layers of the source images are extracted by following equations:

$$D_1 = I_{(x,y)} - B_1 \tag{8}$$

$$D_2 = PAN_{(x,y)} - B_2 \tag{9}$$

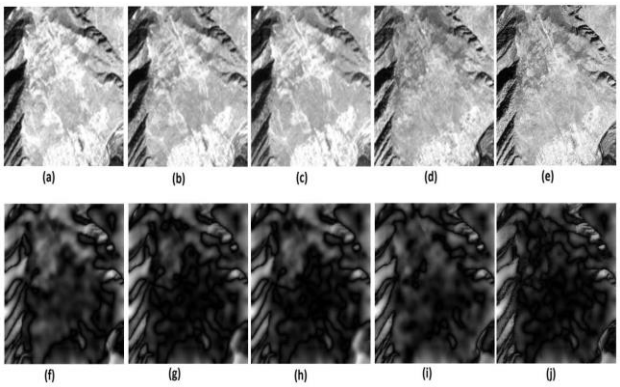


Figure 2. Illustration the saliency maps images of IKONOS data-set, (a)-(e) show the source images of Red, Green, Blue, NIR bands and PAN image, respectively, (f)-(j) show the saliency maps of these images, respectively.

$$S_1 = abs(B_1 - GF_1(I, PAN)) \tag{10}$$

$$S_2 = abs(B_2 - GF_2(PAN, PAN)) \tag{11}$$

where $S_1 and S_2$ are the saliency maps of the source images. For clarity, the input of the guided filter in saliency map S_1 is the PAN image while the band of MS image is considered as a guidance image. In respect to saliency map S_2, the PAN image is considered as the input and guidance image at the same time. The saliency maps for source images are shown in fig. 2.

3.2 The Weight Maps Construction

The special significance of weight maps that to acquire a good fusion result because the MS and PAN images contain information for the same object or scene, therefore, we need to integrate the noteworthy information from the source images. Based on the saliency map which is done in the previous section, we assigning the weight maps for each detail layer which has the tiny details whereas the base layers contain the background information. The weight maps can be extracted by following equations:

$$W_1 = \frac{S_1}{S_1 + S_2} \tag{12}$$

$$W_2 = \frac{S_2}{S_1 + S_2} \tag{13}$$

where W_1 and W_2 are the weight maps of the detail layers of the source images. Fig. 3. Illustrates the weight maps of the source images.

3.4 Base and Detail Layers Fusion

In this section, similarly [14], the base layer fusion is applied by using the mean rule for the two base layers. The following equation is performed the base layer fusion.

$$F_B = \frac{1}{2}(B_1 + B_2) \tag{14}$$

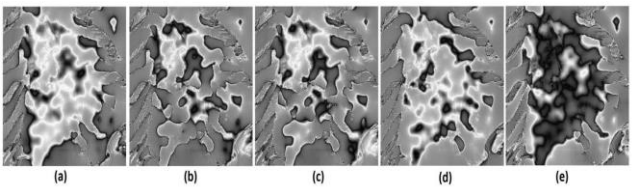


Figure 3. Illustration the weight maps images of IKONOS data-set, (a)-(e) show the weight maps images of Red, Green, Blue, NIR bands and PAN image, respectively.

Here, F_B denotes the fused base layer. In terms of detail layer fusion, is applied by multiplying the detail layer with corresponding the weight map thereafter adding the product of each detail layer. This can be obtained by the following equation.

$$F_D = (W_1 D_1) + (W_2 D_2) \tag{15}$$

where $W_1 and W_2$ are weight maps of $D_1 and D_2$, respectively, and F_D is the fused detail layer. Fig. 4. Shows the fusion result of base and detail layers.

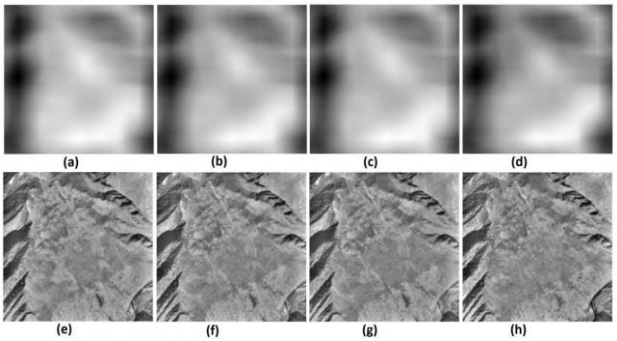


Figure 4. Illustration the fusion result of base and detail layers of IKONOS data-set, (a)-(d) show the base layer fusion results of Red, Green, Blue, NIR bands, respectively, (e)-(h) show the detail layer fusion results of Red, Green, Blue, NIR bands, respectively.

3.5 Final Fusion Result

The final fusion result for each band is applied by adding the fusion result of base and detail layers. After that, the final fusion result for MS image can be obtained by concatenation the fusion result of all bands. The final fusion result for one band can be given by the following equation:

$$F_R = F_B + F_D \quad (16)$$

where $F_R, F_B, and F_D$ represent the final fusion result for one band, the base layer fusion, and detail layer fusion, respectively.

4. THE EXPERIMENTAL RESULT AND EVALUATION

To evaluate the performance of our technique, we are used some real and degraded data sets such as QB, IKONOS, GeoEye. After subjective evaluation should be appraised the performance of each fusion method quantitatively. Hence, six typical evaluation metrics, namely, the correlation coefficient (CC), the root mean square-error (RMSE), universal-image-quality indexes (UIQI), the erreur-relative-global-adimensionnelle de synth'ese (ERGAS), a quaternion based coefficient (Q4) index, and the spectral-angle-mapper (SAM), are used for degraded data while the QNR, D_s, and D_λ are used for evaluation of real data set. Moreover, we compare our fusion result with some existing methods.

Before embarking on experiments, the parameters setting for the guided filter which are the square window r, and ζ are 10 and 0.4^2, respectively. Whereas the square window of the mean filter is 100.

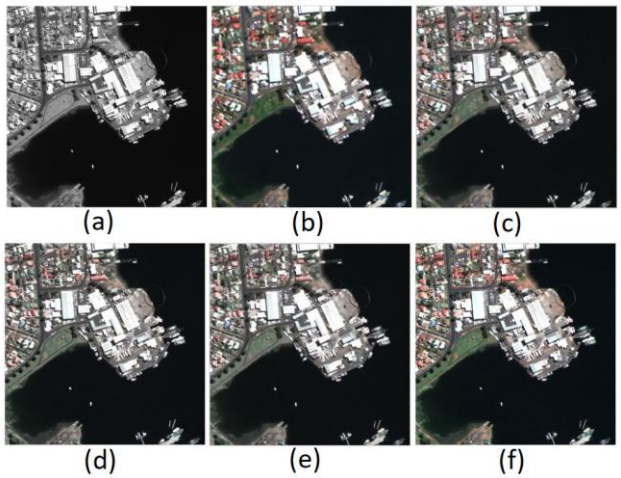


Figure 5. GeoEye images fusion results. (a) PAN image. (b) Up-sampled MS image. (c) proposed method. (d) SSD method. (e) PSD method. (f) GIHS method.

4.1 Experiments with Real Data-Sets

In this section, two full scale data-sets have been using to evaluate the performance of our technique which are QuickBird, and GeoEye. The QuickBird satellite offers PAN and MS images with resolution 0.7m and 2.8m, respectively. The GeoEye provide PAN image of 1m resolution and MS image of 4m resolution. The size of MS image and PAN image in these experiments 256×256, 1024×1024, respectively.

Fig. 5. Shows the fusion results for GeoEye data set of our technique and previous work. Note that, the MS image is up-sampled corresponding with PAN image.

Fig. 5. (a) and (b) show the PAN and up-sampled MS images of the GeoEye data-set. The results of previous work and our technique are shown in Fig. 5 (c)-(f). Visually, the fusion result which is obtained by our work is closed to the up-sampled MS image in the spectral side. In spatial side, all of these methods seem to have been doing well but our work it can be clearly the best one. Table I. Shows the evaluation of spectral and spatial distortions between PAN and MS images and the fused image which are used to compute the QNR. Note that, the best results of every quality indices are clearly noticeable by font **bold**. Here, it can be clearly observed that the proposed method performs the superior results in the spectral and spatial aspects evaluation when compared to the other methods.

Another real QuickBird data-sets, as shown in Fig. 6 (a) and (b), is used to evaluate the performance of our work. The results of the different image fusion methods are shown in Fig. 6 (c)-(f). Visually, Fig. 6 (c) the fusion result which is obtained by our work is closed to the up-sampled MS image in the spectral side. Table 4.2 shows the evaluation of spectral and spatial distortions between PAN and MS images and the fused image which are used to compute the QNR. As we can see from Table II, the biggest value of QNR and the smallest value of D_λ are done by our method. Here, it can be clearly observed that the GIHS method performs the superior results in D_s followed by SSD method.

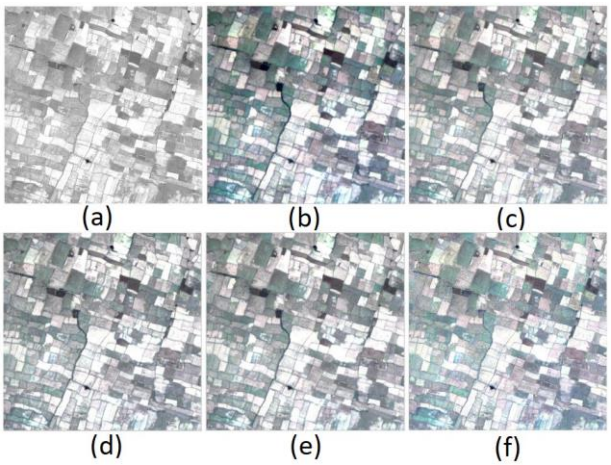


Figure 6. QB images fusion results. (a) PAN image. (b) Up-sampled MS image. (c) proposed method. (d) SSD method. (e) PSD method. (f) GIHS method.

Table 1: Results comparison of the proposed method with some existing approaches on real GeoEye data-sets.

Method	Proposed	SSD	PSD	GIHS
D_s	**0.0087**	0.00871	0.00872	0.0087
D_λ	**0.0981**	0.1903	0.2138	0.1425
QNR	**0.9018**	0.8097	0.7861	0.8574

Table 2. Results comparison of the proposed method with some existing approaches on real QuickBird data-sets.

Method	Proposed	SSD	PSD	GIHS
D_s	0.0038	0.0033	0.0036	**0.003**
D_λ	**0.0178**	0.0326	0.0326	0.0442
QNR	**0.9822**	0.9674	0.9673	0.9558

4.2 Experiments with Degraded Data-Sets

Another group of degraded experimental results of image fusion. The IKONOS system simultaneously offers a four-band MS image with 4m resolution and a single-band PAN image with 1m resolution. The results of the different image fusion methods are

shown in Fig. 7 (c)-(f). The quantitative assessment indices results are shown in Table III. Here, it can be clearly observed that the proposed method performs the superior results in the most evaluation indexes when compared to the other methods.

5. CONCLUSION

A simple pan-sharpening technique using saliency detection based on the guided filter is presented, furthermore, we evaluate our technique by using some real and degraded data-sets then compare our technique with some existing methods in both subjective and objective aspects. By using guided filtering and saliency map in the multi-scale fusion method, the proposed fusion technique accomplishes spatial consistency. The proposed technique has a straightforward execution and is computationally effective. Thus, our work performs the best evaluation among of other existing methods.

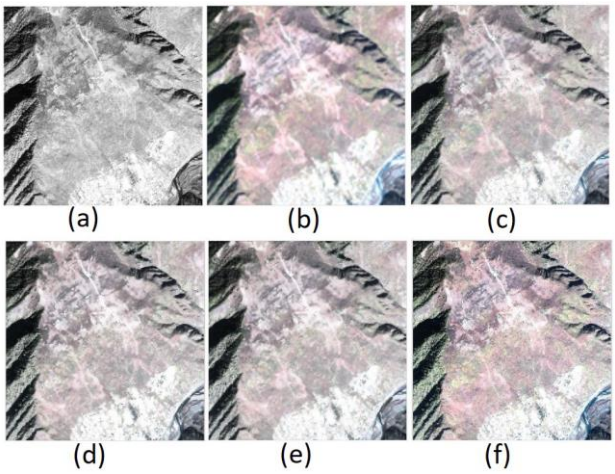


Figure 7. IKONOS images fusion results. (a) PAN image. (b) Up-sampled MS image. (c) proposed method. (d) SSD method. (e) PSD method.(f) GIHS method.

Table 3. Results comparison of the proposed method with some existing approaches on IKONOS data-sets.

Method	Proposed	SSD	PSD	GIHS
CC	**0.9353**	0.9293	0.9331	0.9168
RMSE	**22.35**	23.33	22.66	25.65
ERGAS	**3.26**	3.39	3.30	3.73
SAM	5.9	6.04	**5.88**	6.33
Q4	0.706	0.733	0.724	**0.736**
UIQI	**0.9306**	0.9267	0.9295	0.9133

6. REFERENCES

[1] Aly, H. A., and Sharma, G. (2014). A regularized model-based optimization framework for pan-sharpening. *IEEE Trans. Image Processing*, *23*, 6 (Jun.2014), 2596-2608.

[2] Vakalopoulou, M., Karantzalos, K., Komodakis, N., and Paragios, N. (2016). Graph-based registration, change detection, and classification in very high resolution multitemporal remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*, 7(Jul.2016), 2940-2951.

[3] Dong, W., Li, X. E., Lin, X., and Li, Z. (2014). A bidimensional empirical mode decomposition method for fusion of multispectral and panchromatic remote sensing images. *Remote Sensing*, *6*, 9 (Sep.2014), 8446-8467.

[4] Dalla Mura, M., Prasad, S., Pacifici, F., Gamba, P., Chanussot, J., and Benediktsson, J. A. (2015). Challenges and opportunities of multimodality and data fusion in remote sensing. *Proceedings of the IEEE*, *103*, 9 (Sep.2015), 1585-1601.

[5] Xiao Z, Xie L, Shen L. 2010 Image fusion algorithm based on local correlation in wavelet domain. In *Proceedings of the* Second *International Conference on Internet Multimedia Computing and Service* (Dec 30, 2010) 122-125, ACM.

[6] Liu, C., Jing, Z., Xiao, G., and Yang, B. (2007). Feature-based fusion of infrared and visible dynamic images using target detection. *Chinese optics letters*, *5*, 5 (May.2007), 274-277.

[7] Cremer, F., Schutte, K., Schavemaker, J. G., and den Breejen, E. (2001). A comparison of decision-level sensor-fusion methods for anti-personnel landmine detection. *Information fusion*, *2*, 3 (Sep. 2001), 187-208.

[8] Wang, Xiaobei, N. Rencan, and G. Xiaopeng. (2018). Two-scale Image Fusion of Visible and Infrared Images Using Guided Filter. In Proceedings *of the 7th International Conference on Informatics, Environment*, Energy and Applications, (Mar.2018), 217-221, ACM.

[9] Burt, J. Peter, Adelson, H. Edward. (1987). The Laplacian pyramid as a compact image code. In *Readings in Computer* Vision, 671-679.

[10] Toet, A., Van Ruyven, L. J., and Valeton, J. M. (1989). Merging thermal and visual images by a contrast pyramid. *Optical engineering*, *28*,7 (Jul.1989), 287789.

[11] He, K., Sun, J., and Tang, X. (2013). Guided image filtering. *IEEE transactions on pattern analysis & machine intelligence*, 6, (Jul.2013), 1397-1409.

[12] Toet, A. (2016). Iterative guided image fusion. *PeerJ Computer* Science, 2, e80.

[13] Bavirisetti, D. P., and Dhuli, R. (2016). Two-scale image fusion of visible and infrared images using saliency detection. *Infrared Physics & Technology*, 76 (May.2016), 52-64.

[14] Abugabah, A and Alfarraj, O (2015). Issues to Consider in Designing Health Care Information Systems: A User-centred Design Approach, *electronic Journal of Health Informatics*, Vol 9 (1): e8, pp1-15

[15] Budhiraja, S. (2016). Multiscale image fusion for pansharpening of multispectral images using saliency detection. In *Contemporary Computing (IC3), 2016 Ninth International Conference on, (Aug.2016)* ,1-6. IEEE.

[16] Hu, Y., Xie, X., Ma, W. Y., Chia, L. T., and Rajan, D. (2004). Salient region detection using weighted feature maps based on the human visual attention model. In *Pacific-Rim Conference on Multimedia* (Nov.2004), 993-1000, Springer, Berlin, Heidelberg.

[17] Kou, F., Chen, W., Wen, C., and Li, Z. (2015). Gradient domain guided image filtering. *IEEE Transactions on Image Processing*, 24, 11 (Nov.2015), 4528-4539

Dual-band Welding Speed Monitoring Method Based on Deep Learning

Jionghang Shen, Zhuang Zhao, Jing Han, Yi Zhang
School of Electronic and Optical Engineering, Nanjing University of Science and Technology
Nanjing, China
jionghangshen@qq.com

Lianfa Bai
School of Electronic and Optical Engineering, Nanjing University of Science and Technology
Nanjing, China
blf@njust.edu.cn

ABSTRACT
This paper proposes a welding speed monitoring method based on the weld pool image, which can effectively monitor the welding quality. Through theoretical analysis and experiments, 660 nm band-pass and the 850 nm high-pass are selected as the optimal bands for weld pool image capturing. After capturing dual bands images, through extracting and merging the contours of dual band images, the binary image of weld pool contour is obtained. An improved LeNet deep network is used to process the binary image and train a model to monitor the stability of welding speed. Compared to traditional machine learning method, the improved LeNet deep network has simple structure and obtain high classification accuracy and fast response speed.

CCS Concepts
• Computing methodologies→ Machine learning→ Learning paradigms→ Supervised learning→ Supervised learning by classification.

Keywords
dual-band; weld pool vision; welding speed monitoring; deep learning.

1. INTRODUCTION
Welding is an important processing method in modern industry [1]. With the development of technology, the intelligence of welding has also attracted more and more attention [2]. In the traditional welding process, workers mainly judge the welding quality by the shape of the weld pool and personal experience. The weld pool image contains a large amount of information, which can directly reflect weld quality. It is of important significance in the field of intelligent welding.

The weld pool visual sensing is generally divided into active visual sensing and passive visual sensing [3]. Active visual sensing mainly depends on the external strong light source to inhibit arc interference, with the reflected light of the weld pool as the signal source. Passive visual sensing mainly regards the self-emitted radiation of the weld pool and the reflected arc from the weld pool surface as the signal source.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301537

At present, the main weld pool visual sensing system generally adopts passive visual sensing. When the weld pool is in the process of forming, both the self and the arc light will emit a large amount of light radiation, which will seriously affect the image quality captured by the CCD camera. There are many researches on the passive visual sensing of weld pool. Aiming at the characteristic of aluminum alloy double MIG (Metal Inert Gas) welding, Wang Kehong adopted 980nm band-pass filter and suitable dimming system to obtain a clearer visual image of aluminum alloy double wire weld pool [4]. Wang Jianjun used a secondary filter made of a heat absorbing glass sheet and 640 nm band-pass filter to successfully obtain a clear weld pool image [5]. Yan Zhihong pointed out that the use of 1064nm band-pass filter for weld pool capture can effectively suppress arc interference. However, the filter band is located at the critical point of CCD spectral response and the quantum efficiency is insufficient, which affects the imaging clarity of the weld pool [6]. The above studies show that the choice of filter bands can seriously affect the imaging quality of the weld pool. However, the choice of filter bands is based on the conclusion of experimental results, without systematic and scientific theoretical analysis. The choice of bands proposed by different researchers is quite different and there is no theoretical guidance.

For steel plates of different thickness, the welding speed should be different. On the one hand, too fast welding speed will decrease heat input, resulting in discontinuous weld shape. On the other hands, too slow welding speed will lead to excessive heat input, causing steel plate deformation. There are many researches on welding stability monitoring based on deep learning. These researches are generally based on the geometrical parameters of weld pool image. Asif Iqbal used the ANN (Artificial Neural Network) model to predict the shape of the weld, which proved that the welding speed has a great relationship with the shape of the weld pool and the welded seam [7]. Erwanto used the welding speed, welding current and voltage as the input value of the network, and predicts the width of the weld by convolution neural network [8]. Based on BP (Back Propagation) neural network, Zhang Yunwei used electrical parameters, welding speed and welding current as input parameters to establish a prediction model of weld pool geometric parameters [9]. Ario Sunar Baskoro used binary image of the captured TIG (Tungsten Inert Gas) weld pool image, and treat the pixel width, current value and preset welding speed as input values as input and train a model to predict the welding speed [10]. In general, these methods extract weld pool contour after obtaining the weld pool image. However, these methods are not good for multi-band weld pool images and is easily affected by interference factors such as arcs. Moreover, their network input is generally a preset parameter, so that the

features extracted by the network are limited and often fail to achieve the expected results.

In this paper, a welding speed stability monitoring method is proposed based on the combination of weld pool image and deep learning. Firstly, we proposed how to capture a clear weld pool image with the arc spectrum. Then extract the contour of weld pool through dual-band images. Finally, an improved LeNet deep network is used to monitor the welding speed. The reliability of the method is verified through the CMT+P welding experiments of stainless steel, high strength steel and high nitrogen steel.

2. BAND SELECTION

The imaging system integrates the reflection (or radiation) of the objects through the optical system within its sensitive spectral range, which can be approximately given by the following equation:

$$B=\int_{\lambda_1}^{\lambda_2} L(\lambda)\cdot\rho(\lambda)\cdot\tau(\lambda)\cdot\eta(\lambda)\,d\lambda \qquad (1)$$

B represents the imaging signal strength; $L(\lambda)$ represents the external spectral radiation characteristics; $\rho(\lambda)$ represents the spectral reflectance (or radiation) coefficient of the target; $\eta(\lambda)$ represents the optical system transmittance; λ_1、λ_2 represents the lower limit and upper limit wavelength of the sensitivity range of the imaging device.

For the stainless steel GMAW welding process, there is mainly interference from arc light. These lights will cause strong interference and affect the quality of the captured image. The imaging band should with strong self-radiation and weak arc radiation, and the selected band should be as wide as possible in order to increase the radiation amount of the weld pool. Figure 1 shows the normalized radiation spectrum of the weld pool and arc spectrum normalization. It can be seen that the difference between the two spectrum is larger in the bands under 600-700nm and after 850nm. As a result, this paper uses 660nm band-pass and 850nm high-pass as the capture band of the weld pool.

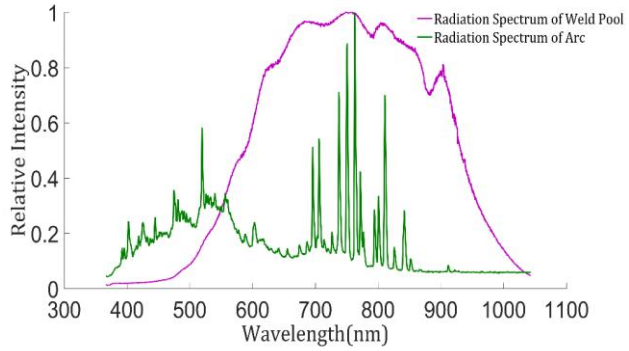
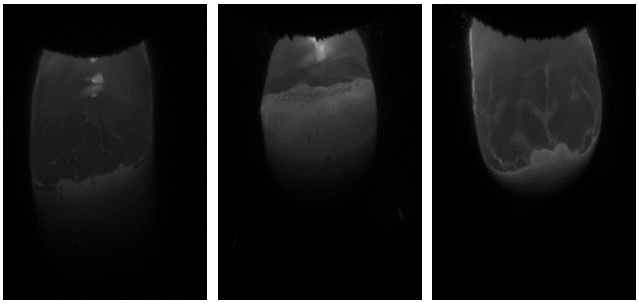


Figure 1. The received arc spectrum normalization curve and the received weld pool self-radiation normalization curve.

3. IMAGE PREPROCESSING

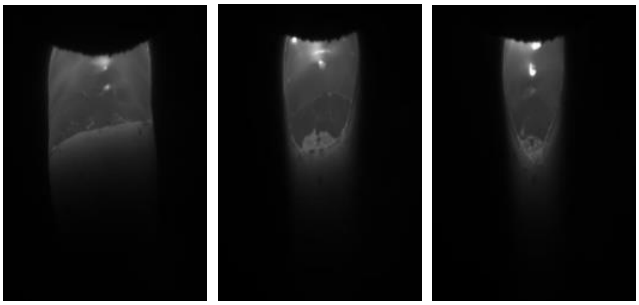
The characteristics of the weld pool with different currents are quite different. This paper mainly analyzes three kinds of welding wires: stainless steel, high strength steel and high nitrogen steel. The weld pool images of these three types welding wire on the stainless steel are shown in Figure 2. As we can see, the weld of stainless steel is more clear than high strength steel and high nitrogen steel which has more complex surface texture in weld pool.

In most cases, the discrimination of different welding speeds on the weld pool image is obvious. Taking stainless steel wire with Ar flow rate 25L/min as an example, as shown in Figure 3, the shape of the weld pool at different welding speeds, especially the width information, is significantly different. In this case, train the image after ROI selection and resize operation can get a good accuracy.



(a) **(b)** **(c)**

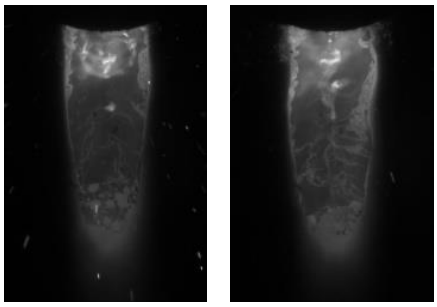
Figure 2. Weld pool image under different welding wires. (a) stainless steel, (b) high strength steel, (c) high nitrogen steel.



(a) **(b)** **(c)**

Figure 3. Weld pool image under different welding speeds. (a) 10cm/min, (b) 30cm/min, (c) 50cm/min.

However, in some cases, especially the weld pool at high currents, the result is quite different. Taking high-nitrogen steel wire with Ar flow rate 25L/min as an example, as shown in Figure 4, the discrimination in the weld pool image obtained at different welding speeds is small, and it is difficult to distinguish through eyes. In this case, the previous operation of the weld pool image cannot get a usable result.



(a) **(b)**

Figure 4. Weld pool image of high nitrogen steel at high current under different welding speeds. (a) 30cm/min, (b) 50cm/min.

In order to monitor the welding speed of this type welding, this paper extract image feature of original image. In terms of welding speed, the visual classification is based on geometric parameters. The welding speed has a great influence on the width, length and aspect ratio of the shape of the weld pool. Thus, we propose to use the contour of the weld pool image to monitor the welding speed.

We used OTSU to segment the images [11]. This method is an efficient image binary method proposed by Japanese scholar Otsu. The binary threshold selection criterion is to maximize the inter-class variance of the target and background.

The segmentation threshold of foreground and background is t. w_0 represents the ratio of former attractions points to the image. μ_0 represents the average gray scale. w_1 represents the ratio of background point to the image. μ_1 represents the average gray scale. The variance between classes is g.

$$g = w_0 * w_1(\mu_0 - \mu_1)^2 \quad (2)$$

The threshold T which maximizes the variance between classes can be obtained by the traversal method. After eroding and dilating to remove small contours, we will get the final weld pool contours.

In practical applications, as shown in Figure 6, for the 660 nm band-pass filter image, the grayscale value of the weld pool head image is larger, and the difference between the tail and head image of the weld pool is obvious, so the extracted contour has a large defect at the tail. Meanwhile, due to the existence of arc, the head contour of the weld pool is often affected. The resulting contour image is shown as Figure 7(a). For the 850nm high-pass filter image whose resulting contour image is shown as Figure 7(b), the difference between the tail image of the weld pool and the background area is not obvious, so there is a certain defect in the extraction of the tail contour of the weld pool.

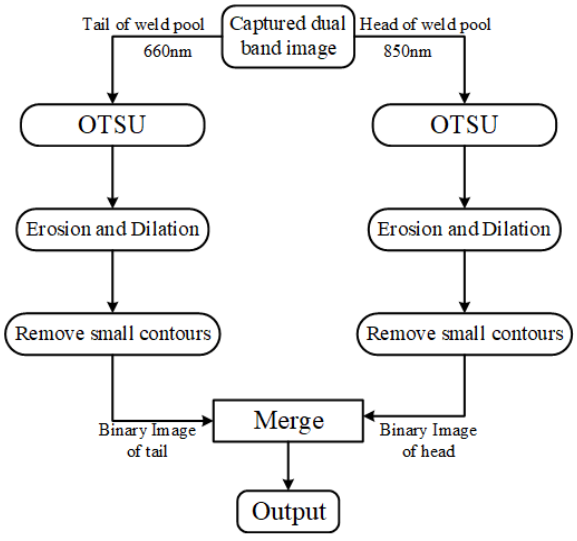
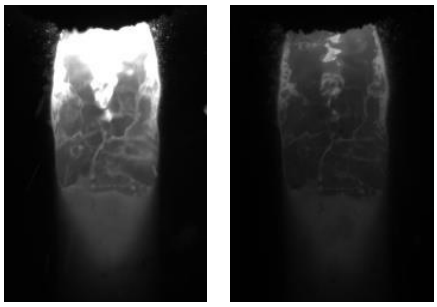


Figure 5. 660 band-pass and 850 high-pass filter image processing flow.

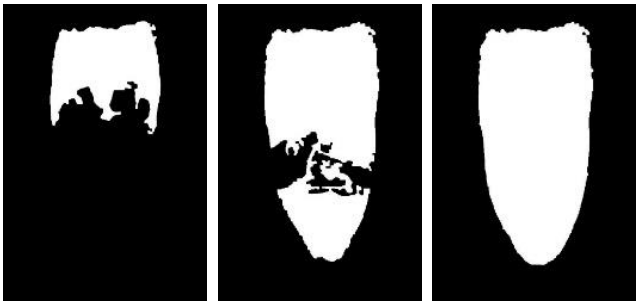
This paper comprehensively analyzes the characteristics of the two-band weld pool image. The 850nm band image can clearly obtain the head contour of weld pool. And when we see the contour extracted from the image in the 660 nm band in Figure 7(a), we feel that removing the brightest part of the weld pool could extract a complete weld pool tail contour. As a result, as shown in Figure 5, the 660nm band-pass filtered image and the 850nm high-pass filtered image are processed by a series of processing using OTSU threshold method. The 660 nm band-pass filter image is used to extract the tail contour of the weld pool, and the 850 nm high-pass filter image is used to extract the head contour of the weld pool. Finally, the two are merged through the binary image to obtain the overall contour of the weld pool.

Taking high-nitrogen steel wire, welding current 170A, welding speed 24cm/min as an example, the original picture is shown in Figure 6, the binary image extracted by a single weld pool image and the weld pool binary image extracted by this method are shown in Figure 7. As shown in Figure 7, the proposed method of this paper can extract the contours of the weld pool shape more clearly.



(a) **(b)**

Figure 6. Weld pool image of 660nm band-pass and 850nm high-pass. (a) 660nm band-pass, (b) 850nm high-pass.



(a) **(b)** **(c)**

Figure 7. Contour image extracted by different methods. (a) 660nm band-pass contour image (b) 660nm high-pass contour image (c) contour image extracted by this paper's method.

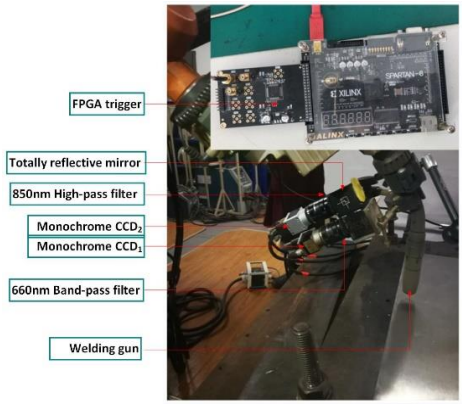


Figure 8. Experimental system.

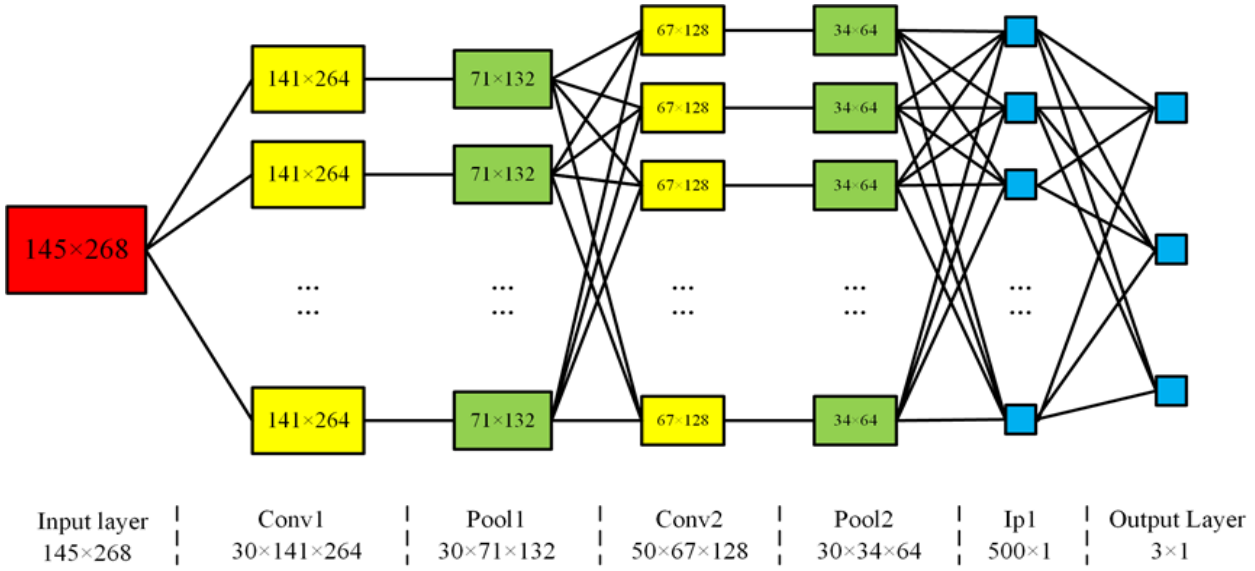


Figure 9. The network structure diagram used in this paper.

4. EXPERIMENT

In order to obtain the weld pool image during the welding process, we designed the experiment system as show in Figure 8. The experiment system includes one welding gun and two CCD cameras with different filters which is triggered by FPGA. To classify the welding speed based on the weld pool image, we need to choose an appropriate network. We choose the improved LeNet-5 convolution network as the classification network of weld pool image [12]. The LeNet-5 network is a convolutional neural network proposed by Professor Yann LeCun for handwriting recognition. It consists of 7 layers, including two convolutional layers, two pooling layers and three full connect layers. Its structure is shown in Figure 9.

4.1 Data preparation and model training

Most area of the original image captured by the system is black. It has no meaning on image classification. Therefore, we select a dynamic ROI that includes the entire weld pool based on the extracted contour.

Depending on the welding speed, the number of weld pool images captured during each welding process varies from 300 to 500 pictures. We use 7 sets of data as the training set, which contains about 2500 pictures, and the other three sets of data which contains about 1200 pictures as the validation set and test set. According to the welding speeds, we divide the pictures into three categories, and the labels of each picture are set to 0, 1, and 2(low, medium, high).

For training, the ROI area image is still too large, so all the pictures were resized to 145×268 size when we made the lmdb data set. This can reduce hardware requirements and training time during training.

After generating the lmdb data set, we adjust the network parameters. We set the net output number to 3, the learning rate to 1×e-8, and the number of iterations to 20000.In the end, we start training.

4.2 Result

Depending on the welding current and the welding wire, we train different models separately. We use the original and contour maps to train in the deep network, and compare them with the traditional contour-based KNN algorithm [13][14]. After the training, we get the classification accuracy. The experimental results are shown in the table 1.

It can be seen that the accuracy of the classification results based on the original image and the contour image is close to 1 on the stainless steel welding wire with obvious difference in the weld pool image of each welding speed, which is higher than the accuracy of the traditional KNN algorithm. On the high strength steel and high nitrogen steel welding wire, the results are very different. At low currents, the classification results based on the original image and the contour image are close to 1. At medium and high currents, the classification accuracy based on the original image is significantly lower due to the less difference between the weld pool images at different welding speeds. However, compared with the deep learning based on the original image and the traditional machine learning classification algorithm, the classification results based on the contour map used in this paper has a higher accuracy, though the accuracy of the stainless steel welding wire has decreased.

Table1. Classification results of weld pool image

Welding wire	**Stainless steel**			**High strength steel**			**High nitrogen steel**		
Current value	**80A**	**120A**	**160A**	**101A**	**130A**	**174A**	**101A**	**130A**	**174A**
LeNet(original)	99.8	100	100	98.9	85.8	81.0	97.8	97.8	64.4
LeNet(contour)	**99.9**	**100**	**100**	**98.3**	**89.9**	**93.8**	**99.9**	**98.3**	**84.2**
KNN(contour)	98.9	98.9	93.3	97.3	85.5	74.5	97.8	96.2	72.9

The network model we obtained can monitor the welding speed online during welding process, and can give feedback to the operator when the welding speed is too high or too low.

5. CONCLUSION

In this paper, a passive dual path weld pool vision sensing system is established. We confirm the 660nm band-pass and 850nm high-pass are the best bands to capture the image of the weld pool.

This paper proposes a method for extracting the contour of the weld pool based on the weld pool images under dual bands. The OTSU threshold segmentation method is used to extract the head part contour of weld pool 660nm band and tail part of weld pool image on 850nm band, and finally merge to obtain the accurate weld pool contour.

Based on the obtained clear pool contour image, a effective method based on deep learning is proposed to classify the different welding speed and obtain the model for monitoring welding stability.

6. ACKNOWLEDGMENTS

"This work was supported by the National Natural Science Foundation of China (61727802)，National Defense Science and Technology Innovation Project(17***01)"

7. REFERENCES

[1] Zhang, G., and Yongzhe, L. I. 2016. Towards intelligent welding in the context of industry 4.0. *Aeronautical Manufacturing Technology*, 59, 11 (Nov. 2016), 28-33.

[2] Chen, S. B., Tao, L., Jie, C. W., and Tao, Q. 2004. Concepts and technologies on intelligentized welding manufacturing engineering. *Transactions of the China Welding Institution*, 25, 6 (June. 2004), 124-128.

[3] Chen, F. 2009. *Research on Characteristic of Divisional Weld Pool Visual Image in Spray Transfer Welding*. Ph.D. Dissertation. Nanjing University of Science & Technology, Nanjing, China.

[4] Wang, K. H., Yang, J., Feng, Q., and Shen, Y. J. 2007. Molten pool image gathering and processing of aluminum alloy twin-wire mig welding. *Transactions of the China Welding Institution*, 28, 1 (Jan. 2007), 53-56+60.

[5] Wang, J. 2003. Image taking and processing of molten pool characters during aluminum alloy tig welding. *Chinese Journal of Mechanical Engineering*, 39, 5 (May. 2003), 125-129.

[6] Yan, Z. H., Zhang, G. J., Qiu, M. Z., Gao, H. M., and Wu, L. 2005. Monitoring and processing of weld pool images in pulsed gas metal arc welding. *Transactions of the China Welding Institution*, 26, 2 (Feb. 2003), 37-40.

[7] Iqbal, A., Khan, S. M., and Mukhtar, H. S. 2011. ANN assisted prediction of weld bead geometry in gas tungsten arc welding of HSLA steels. *In Proceedings of the World congress on engineering 2011 Vol I*. WCE, London, UK, 6-8.

[8] Baskoro, A. S. 2011. Monitoring of molten pool image during pipe welding in gas metal arc welding (GMAW) using machine vision. *In Advanced Computer Science and Information System (ICACSIS), 2011 International Conference on. IEEE*. IEEE, Piscataway, NJ, 381-384.

[9] Zhang, Y. W. 2012. *Vision-based Sensing of the Topside Weld* Pool *Geometry in PAW*. Ph.D. Dissertation. Shandong University, Jinan, China.

[10] Baskoro, A. S., Rahman, A. Z.,and Haikal. 2017. Automatic welding speed control by monitoring image of weld pool using vision sensor. *ARPN Journal of Engineering and Applied Sciences*, 12, 4 (Apr. 2017), 1052-1056.

[11] Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9, 1 (Jan. 1979), 62-66.

[12] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 11 (Nov. 1998), 2278-2324.

[13] Cover, T., and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13, 1 (Jan. 1967), 21-27.

[14] Mejdoub, M., and Amar, C. B. 2013. Classification improvement of local feature vectors over the KNN algorithm. *Multimedia Tools and Applications*, 64, 1 (May. 2013), 197-218.

Template Attentional Siamese Network for Object Tracking

Junyan Gao
Guangdong University of Technology
Guangzhou, China
+86 17724223750
junyangaogdut@gmail.com

Zhenguo Yang*
Guangdong University of Technology
Guangzhou, China
+86 18320145447
yzgcityu@gmail.com

Wenyin Liu*
Guangdong University of Technology
Guangzhou, China
+86 13910586417
liuwy@gdut.edu.cn

ABSTRACT
Recent years, visual object tracking has attracted more and more attention as a fundamental topic. Many deep based trackers, especially Siamese Network based trackers, have achieved state-of-the-art performance on multiple benchmarks. However, most of these trackers applied with the first frame as template throughout the tracking process. We propose a Template Attentional Siamese Network called TASNet. The core of TASNet is combining the detection results of two template frames, where the first frame extracting discriminative features and the latest frame capturing the motion changes, to enhance model tracking effect. Template-wise weights are calculated from attention mechanism to integrate the detecting results of two templates in current frame tracking. The proposed architecture is trained from end to end on the ILSVRC2015 video dataset. Our tracker operates at frame-rates real-time and achieves state-of-the-art tracking accuracy while large deformation of the object is appeared.

CCS Concepts
• **Computing methodologies → Tracking**

Keywords
attention mechanism; discriminative features; motion change; object tracking; Siamese network

1. INTRODUCTION
Visual object tracking is a basic and challenging task in computer vision. Given the bounding box of the target in the first frame, the aim is to locate the target in all the following frames in a video sequence. It plays vital application values in video surveillance, video conferencing, automatic driving, and human-computer interaction. However, it still remains very great challenging in large appearance variance like illumination, scale variation, deformation, occlusions and fast motion. Besides, the speed and high accuracy are also required by practical applications.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301544

Existing trackers can be roughly divided into two branches. One of the branch is Correlation Filter (CF), which trains a correlation filter classifier in Fourier domain. One notable example is Kernel Correlation Filter (KCF) with the running speed of about 172 frames per second used HOG feature. The CF algorithm achieves high running speed for the replacement of the exhausted convolution with element-wise multiplications using Fast Fourier Transform. However, the tracking accuracy of this branch is not high enough to practical applications. The other of the branch is deep learning model, which learns very strong deep features through multiple convolution neural layers. The deep learning models are trained aims to improve the tracking accuracy due to their strong feature learning and representation ability. By extensively training deep learning network on large datasets offline, the model achieves far better performance than CF algorithm. However, no updating the template frame during tracking process limits the model to attain consummate results.

In this paper, we show the deep learning tracker, which offline trained and online updated the template frame. In Fig. 1, the proposed Template Attentional Siamese Network (TASNet) consists of a template attention branch and two detection branches. The main contributions of our work can be summarized as follow:

1) An end-to-end deep learning tracker is designed for object tracking, especially dealing with the situation of large deformation and template drift.

2) The template attention mechanism is explored within the TASNet for better integration of the detection results of the two templates.

3) We achieve competitive results compared with the state-of-the-art trackers in OTB50 [1], OTB100 [2], and VOT2017 [3] with the real-time speed of 40 FPS, which proves its advantages in both accuracy and efficiency.

2. RELATED WORK
Deep feature based trackers. Recently, deep learning has been widely employed to improve the performance of object tracking due to its strong feature representation power. DLT [4] trained a stacked denoising autoencoder offline to learn general representation of image, and then tracked the object with online fine-tune. It is the first tracking model to apply deep feature to single-target tracking tasks. HCF [5] and HDT [6] exploited features extracted from different layer of deep convolutional neural networks. C-COT [7] and ECO [8] proposed a multi-scale adaptation solution based on the continuous convolution filters. MDNet [9] trained a end-to-end CNN framework, which learning multi-domain generic target representation for single object

tracking tasks. TCNN [10] maintains multiple CNNs in a tree structure to represent target appearances at different states. However, speed and precision have always been difficult to balance.

Siamese network based trackers. Siamese network regards visual object tracking task as a deep similarity learning problem. Siamese network consists of two branches: one branch for extracting features of template image, the other branch for extracting features of candidate image. By comparing the template patch features with the candidate patch features in a search region, the object location is tracked where the highest similarity score is obtained. GOTURN [11] regarded object tracking tasks as a regression problem. It trained deep regression networks offline with labeled video and predicted bounding box online at a very fast speed 100 fps. SiamFC [12] trained a end-to-end fully-convolutional siamese network framework for real-time object tracking in video. SA-Siam [13] introduced a semantic branch to extract semantic features of the image. It improved the tracking performance by combining the appearance branch and semantic branch. Siamese-RPN [14] added region proposal subnetwork after Siamese subnetwork to compute the location and scale of the target precisely. However, using only the first frame as a template limits the increase in accuracy, although it can achieve very fast speed on tracking online.

Attention mechanisms. In recent years, attention mechanisms have been widely used in various fields of deep learning, such as image processing, speech recognition, and natural language processing (NLP). Understanding how the attention mechanism works and applying it to these tasks are great necessity for researchers. SA-Siam [13] proposed channel attention mechanism which assigned a high weight to the activated channel. The algorithm calculated different weight for different channel due to different target activate different feature channel. FlowTrack [15] made use of the rich flow information with spatial-temporal attention to make up for the lack of appearance features of current frame. The algorithm benefited from motion and inter-frame information. RASNet [16] introduced three kinds of attention mechanisms to improve the discriminative ability of the model: general attention, channel favored feature attention and residual attention. The core of the attention mechanism is to learn the most critical information from a wide range of information. In this paper, we proposed a attention mechanism to learn the weights of templates through an end-to-end network.

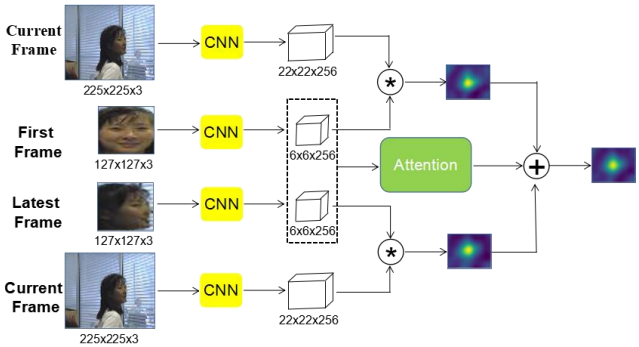


Figure 1. The architecture of the proposed Template Attention Siamese Network (TASNet).

3. PROPOSED TRACKING FRAMEWORK

In this section, we describe the proposed Template Attentional Siamese Network (TASNet) in detail. As shown in Fig.1, the proposed framework consists of two detection branches (the first frame as template and the latest frame as template) and a template attention branch.

3.1 Siamese Network with Dynamic Tracking

For object tracking task, the ground truth in the first frame can usually clearly distinguish the target, such as the positive face, the head of car, the rear of car. There is reasonable evidence to use the first frame as a template in Siamese Network due to its discriminative features. However, using only the first frame as a template limits the improvement of tracking performance because of the dynamic variability of the target in the video. It always appears that the target in current frame is little or no similar to it in first frame. At this point, the latest frame or the latest continuous frames can successfully capture the motion changes of the target [17].

The fundamental idea behind the proposed framework is considering the first frame as template and the latest frame as template jointly. Therefore, the long-term dynamic visual tracking can be achieved even large deformation and occlusion appears.

The input of the network is three image patches: the target patch cropped from the first frame, the target patch cropped from the latest frame (called the t-1 frame), and the current frame (called the t frame). We use notations x_{first}, x_{latest}, z to denote the inputs respectively shown in Fig 1. Both x_{first} and x_{latest} have a size of 127×127×3. The current frame has a size of 255×255×3. The two detection branches take (x_{first}, z), (x_{latest}, z) as input respectively. The convolutional network used to extract appearance features clones the SiamFC network [12], and the features extracted are denoted by $\varphi(x_{first})$, $\varphi(x_{latest})$ and $\varphi(z)$. The response map from the detection branches can be written as :

$$f(z, x_{first}) = corr(\varphi(z), \varphi(x_{first})) \quad (1)$$

$$f(z, x_{latest}) = corr(\varphi(z), \varphi(x_{latest})) \quad (2)$$

Where $corr(\bullet)$ is the correlation operation.

3.2 Template Attention Architecture

We consider the two most representative frames as templates jointly. However, the difference in the target movement changes results in different reference weights for the two templates. This leads to a question: how to determine the reference weights for the two templates? To solve this problem, encouraged by [13] and [16], we design a template attention module to dynamically learn the weights.

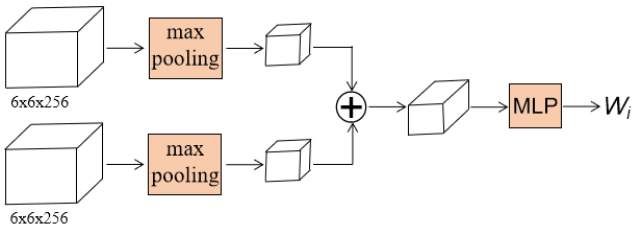


Figure 2. Template attention generates the weight through max pooling and multilayer perception machine. The weight

coefficient W_i includes $W_{x_{first}}$ for template x_{first} and $W_{x_{latest}}$ for template x_{latest}.

Fig. 2 shows the architecture of template attention module. The module takes $(\varphi(x_{first}), \varphi(x_{latest}))$ as input. The following max pooling operation is performed to keep the most important feature information and reduce the amount of calculation. Then a two-layer multilayer perceptron (MLP) and a Sigmoid function are adopted to produce the weights W_i for two templates.

In the end, the final response map is computed as the weighted average of the two response maps from two detection branches:

$$F(z, x_{first}, x_{latest}) = W_{x_{first}} f(z, x_{first}) + W_{x_{latest}} f(z, x_{latest}) \quad (3)$$

Note that we learn the attention parameters in training phase, while using with no updating in testing phase.

4. EXPERIMENTS

In this section, we present the implementation details, evaluation methods and experiments results compared with several state-of-the-art trackers on OTB50, OTB100, and VOT2017 benchmarks.

4.1 Implementation Details

Our framework is pre-trained offline on the ILSVRC2015 [18] video dataset with TensorFlow. We keep the convolutional network used for extracting appearance features and its parameters same as [12], and train attention module parameters to explore optimal weight for two templates. Note that we extract appearance features of the current frame only once in tracking phase.

Our experiments are produced on a machine equipped with an Intel Core i7-6850K 3.60GHz CPU, and a NVIDIA TITAN X (Pascal) GPU. Conventionally, a tracking speed beyond 25fps is considered real-time. Our online tracking speed average at 40 FPS. During evaluation and testing, three scales are searched to handle scale variations.

4.2 Comparison Results

Experiments on OTB50 and OTB100 benchmarks. OTB50 and OTB100 are widely used public tracking benchmarks. OTB50 consists of 50 fully annotated sequences. OTB100 expands the OTB50 to 100 object tracking sequences. We consider two evaluation methods on OTB benchmarks: the precision at different location error threshold and the success rate at different overlap threshold. We compare our framework TASNet with ECOhc, SiamFC, CFNet, STAPLE [19] on OTB50/100 benchmarks. And the precision plots and success plots are shown in Fig. 3. The comparison shows that TASNet achieve the best performance among these real-time trackers on OTB50/100 benchmarks. Our model is an improvement on the basis of SiamFC, so we do several comparative experiments. Fig. 4 shows several comparison results on several OTB100 sequences.

Experiments on VOT2017 benchmark. Visual-Object-Tracking Challenge (VOT) is the most authoritative evaluation platform for international target tracking. The sequences of the VOT dataset are updated every year, and the accuracy of the annotations is improved year by year. We select Accuracy (called A), Robustness (called R), and Expect Average Overlap Rate (called EAO) as evaluation criteria. Table 1 shows the comparison results of our framework TASNet with C-COT, ECOhc, SiamFC, Staple on VOT2017 benchmark. The results prove that our framework achieves competitive performance.

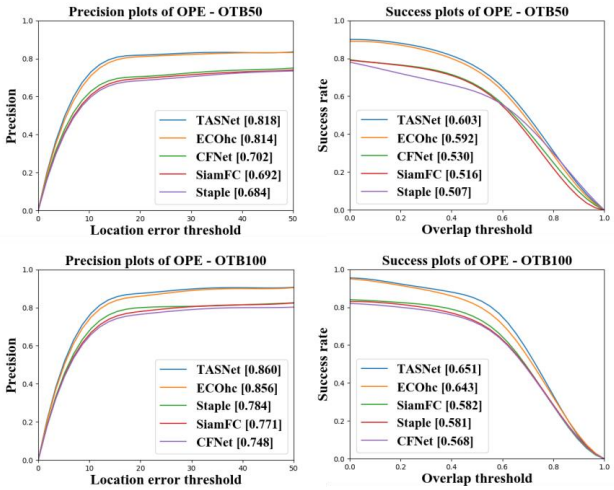


Figure 3. The precision plots and success plots on OTB50/100 benchmarks.

Table 1. Comparison results on VOT2017 benchmark.

Tracker	A	R	EAO
C-COT	0.486	0.432	0.233
ECOhc	0.494	0.435	0.238
SiamFC	0.502	0.585	0.188
Staple	**0.530**	0.688	0.169
TASNet	0.496	**0.386**	**0.257**

5. CONCLUSION

In this work, we present the Template Attentional Siamese Network (TASNet) which is end-to-end offline trained and online tracked. We select the first frame and the latest frame as template jointly, and make use of attention mechanism to compute optimal weight for them to achieve better tracking performance. As a result, the proposed TASNet outperforms other real-time trackers on the OTB50/100 and VOT2017 benchmarks. In the future, we plan to continue exploring the effective fusion ways of different template patch in object tracking task.

6. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No.61703109, No.91748107), China Postdoctoral Science Foundation (No.2018M643024), and the Guangdong Innovative Research Team Program (No.2014ZT05G157).

7. REFERENCES

[1] Wu, Y., Lim, J., and Yang, M.-H. 2013. Online Object Tracking: A Benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2411 – 2418.

[2] Wu, Y., Lim, J., andYang, M.-H. 2015. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9):1834 – 1848.

[3] Kristan, M., and et al. 2017. The visual object tracking vot2015 challenge results. In Proceedings of the IEEE international conference on computer vision workshops.

[4] Wang, N., and Yeung, D. 2013. Learning a deep compact image representation for visual tracking. in Advances in neural information processing systems.

[5] Ma, C., Huang, J.-B., Yang, X., Yang, M.-H. 2015. Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision

[6] Qi, Y., Zhang, S., Qin, L., Yao, H., Huang, Q. Lim, J., Yang, M.-H. 2016. Hedged Deep Tracking. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

[7] Danelljan, M., Robinson, A., Khan, F. S., Felsberg, M. 2016. Beyond correlation filters: Learning continuous convolution operators for visual tracking. in Proceedings of European Conference on Computer Vision.

[8] Danelljan, M., Bhat, G., Khan, F. S., Felsberg, M. 2016. ECO: Efficient Convolution Operators for Tracking. arXiv preprint arXiv:1611.09224.

[9] Nam, H., Han, B. 2016. Learning Multi-domain Convolutional Neural Networks for Visual Tracking. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 4293-4302.

[10] Nam, H., Baek, M., Han, B. 2016. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking. arXiv preprint arXiv:1608.07242.

[11] Held, D., Thrun, S., and Savarese, S. 2016. Learning to Track at 100 FPS with Deep Regression Networks. In Proceedings of European Conference on Computer Vision, pages 749-765.

[12] Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., Torr, P. H. S. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of European Conference on Computer Vision, pages 850-865.

[13] He, A., Luo, C., Tian, X., Zeng, W. 2018. A Twofold Siamese Network for Real-Time Object Tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

[14] Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X. 2018. High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

[15] Zhu, Z., Wu, W., Zou, W., Yan, J. 2018. End-to-end Flow Correlation Tracking with Spatial-temporal Attention. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

[16] Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., Maybank, S. 2018. Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

[17] Guo, Q., Feng, W., Zhou, C., Huang, R., Wan, L., Wang, S. 2017. Learning Dynamic Siamese Network for Visual Object Tracking. In Proceedings of IEEE International Conference on Computer Vision.

[18] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211-252.

[19] Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P. H. S. 2016. Staple: Complementary Learners for Real-Time Tracking. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Figure 4. Several tracking results on OTB100 sequences. The red box is drawn from our framework TASNet, while the yellow box is drawn from SiamFC.

Enhanced Satellite Imaging Algorithm Classifier using Convolutional Neural modified Support Vector Machine (CNMSVM)

Edgar Bryan B. Nicart
Graduate Studies Program
Technological Institute of the Philippines, QC
Camarines Norte State College
Daet, Camarines Norte
edgarbryann@gmail.com

Tony Y.T. Chan, Ph. D.
Graduate Studies Program
Technological Institute of the Philippines
Quezon City Philippines
TIPTonyChan@gmail.com

Ruji P. Medina, Ph. D.
Graduate Studies Program
Technological Institute of the Philippines
Quezon City Philippines
ruji_p_medina@yahoo.com

ABSTRACT
The development of Enhanced Satellite Imaging Algorithm is a continuous effort to achieve further advancement in the area of image processing and recognition through the use of artificial intelligence theories and applications. This study utilized a convolutional neural modified support vector machine aimed at satellite image map classification via new model. The dataset comes from Kaggle using satellite images of the Amazon Rainforest to train multilabel classifier. The result was evaluated using F_2 score which achieved at least 0.91412 by training CNMSVM proposed model.

CCS Concepts
• **Computing methodologies---Modeling and simulation---Simulation types and techniques---Massively parallel and high-performance simulations.**

Keywords
Artificial intelligence, convolutional neural network, satellite image, support vector machine.

1. INTRODUCTION
Satellite image is one of the most powerful and important tools used nowadays and its application ranges from meteorology, oceanography, fishing, agriculture, biodiversity, conservation, forestry, landscape, geology, cartography, regional visible color planning, education, intelligence, and warfare [10]. Interpretation and analysis of satellite imagery however, are conducted using specialized techniques and applications such as using machine learning for effective image classifications. As greatly expressed by famous scientists, there is a need for a model with large learning capacity just to be able to learn thousands of objects from the million images that are of prime importance [11]. Hence, this paper is basing its contribution on three aspects as follows: 1) improve satellite image map classification; 2) promote classification techniques using deep learning while solving the problem of data imbalances; and 3) provide an enhanced image classification function for large datasets. The importance of image classification is largely contributed to the growing demand for advancement in image processing technology enabling different tools and advancement introducing various innovations through the application of different algorithms and machine learning representations. One of such algorithm making a popular trend nowadays is the Convolutional Neural Network (CNN) which has shown strength in its capability to discover relevant contextual features in image categorization problems other than being a deep learning network [1]. This deep learning method has been used to lessen the error on object detection and segmentation and has been used in almost all the application it is believed to be useful such as in detecting vehicles using satellite images [2] pedestrian detection in robotics and autonomous cars [3] image processing application on the depth of lights [4] are among such use.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301545

However, a sensitive part of CNNs is the amount of training data required to properly learn network parameters and also, the power of CNNs to take large context to conduct predictions comes at a price of losing resolution for the output leading to coarse resolution. Therefore, the use of support vector machine to augment this issue by refining and improving classification rates is one of the strategy this study would like to prove. Fine tuning pre-trained CNNs has shown to produce high performance in a short period of time based on some studies [5].

Although CNNs are performers when it comes to classification of maps using multisource satellite imagery, but still it provided misclassifications especially on small objects [6]. On the other hand, multiple-kernel SVM alone is not suited for large-scale data because it needs more training time than SVM. Also, there is the problem of data imbalances which affects most of the supervised learning tasks [7]. Thus, the proposed Convolutional neural modified support vector machines (CNMSVM) is expected to improve this situation.

Given the above scenario, the advent of Satellites and remote sensing makes it possible to detect changes in the earth's surface through monitoring the local, regional and global environmental resources from space. This technological development is made possible through new sensors and higher spatial and spectral resolutions and images being available from satellites over shorter time intervals than ever before [8]. The use of remotely sensed images is becoming a common approach to determine environmental patterns and impact of human activities. Using this process, temporal difference can be discovered requiring adequate

processing of multi-temporal data, proper geo-referencing, and classification. The use therefore provides huge opportunities for monitoring landscapes especially forests which is becoming more and more in danger of being deforested hence, satellite images can give significant analysis and insights on the cause and effect of deforestation and be able to respond properly given the right tool and technology such as the introduction of new found algorithms such as machine learning [9].

2. RELATED WORK

The use of deep learning in machine learning has provided progress especially in the advent of big data models and new requirements for high performance data analytics. Also, right now a research in diverse types of deep learning's networks is being developed such as deep convolutional nets which have brought breakthrough in processing images and speech. [12]

Convolutional neural networks (CNNs) is one of these vision applications rendered for deep learning which is also good for cluttered environment which is a key ingredient for a performance that is compositional in nature of the representations that are learned. However, this deep learning application makes miscalculations caused by differences of color contrast among image tiles that were patches of satellite images from heterogeneous sources [13]. Therefore, it is also important that images taken should be also pre-processed to adjust color contrasts among image tiles acquired so an algorithm may be provided beforehand to address this problem. This enables the images to improve better accuracy in terms of classification and define better the pixels making it easier to calculate thereby completing the training tasks garnering better output. Another important consideration with complicated process as CNNs is the limitation of CPU-based processing that is why among the cited studies it is always pointing out to better utilize GPU to build a better classifier and computing power.

However, no matter how popular convolutional neural networks (CNNs) are it is poor at localizing objects precisely. There are two reasons for coarseness in CNN classification: 1) there is a structural limitation of CNN to carry out fine-grained classification and 2) lack of spatially accurate reference data for training affects CNN judgment. [1] Moreover, there are still recent approaches that aim to overcome the structural issues that lead to coarse classification maps.

Meanwhile, a hybrid clustering algorithm and feed-forward network classifier for land-cover mapping of trees, shade, building and road was explored in the study of [14]. Their study emphasized pre-processing procedure for image to become suitable for segmentation. This optimization algorithm concept provides an idea about improving segmentation in satellite images using ANN in the form of a feed-forward neural network which can primarily increase the chance of the present study to understand improvement of segmentation.

3. SYSTEM ARCHITECTURE/DESIGN

3.1 Support Vector Machine (SVM Classifier)

The SVM, as a new effective statistical learning method, has the superiority compared with traditional classifiers in image classification. SVM is built on the principle of structural risk minimization (SRM). So we need to find the separating hyperplane with a maximum margin $w \bullet x+b=0$, which can satisfy the decision function $f(x,\alpha)$ of SRM.

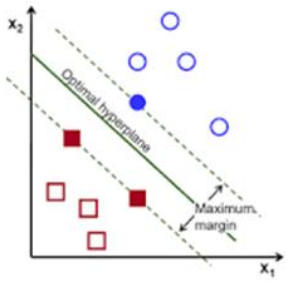


Figure 1. Support Vector Machine architecture visual representation.

3.2 Convolutional Neural Network

The study utilized a total of eleven fully connected layer of a convolutional neural network since it will be dealing with big data higher processing power were considered. To make the processing faster, a reduction of image was done on a 32 x 32 x 3 and the mean image was subtracted to ensure zero center. The so called rectified linear unit (ReLu) activation function for each layer was introduced for efficient computation. Batch normalization with a dropout rate of 0.5 was considered and Adam optimizer for gradient update was applied.

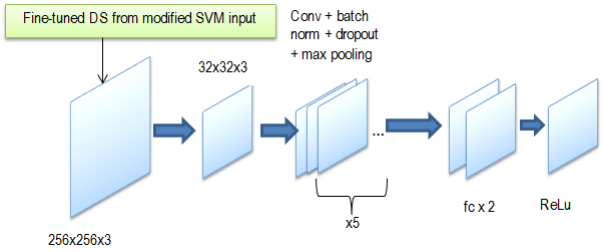


Figure 2. Convolutional neural network with image resizing and trained datasets.

The CNN shows an approach where each neuron θ_n is connected to a local region of the image called receptive field. The calculated weight matrices W_n for each neuron are shared for the whole image. Each neuron θ is considered as a filter with the size of the receptive field to compute the convolution. The filter mask corresponds to the calculated weight matrix which slides over the image. The results are n feature maps where n is the depth of the given data volume.

4. METHODOLOGY

The experiment provided pre-trained models using the modified SVM before using as an input to the deep convolutional network architecture. For each network the weighted sigmoid cross entropy loss was chosen because the F_2 score evaluation result favors recall thereby penalizing false negatives and improve F_2 score.

In addition, separate classifiers are trained for atmospheric labels and land labels since most if not all of the images only had one atmospheric label along with zero or more land labels. By using a Softmax classifier for the atmospheric labels and a sigmoid classifier for the land labels, it was attempted to ensure that each image had only a single atmospheric label.

4.1 Datasets and Classifications

The satellite image taken from Kaggle has a ground-sample distance (GSD) of 3.7m and an ortorectified pixel size of 3m. The data comes from Flock 2 satellites in both sun-synchronous and International Space Station (ISS) orbits and was collected between January 1, 2016 and February 1, 2017. The images were from the Amazon basin from Brazil, Peru, Uruguay, Colombia, Venezuela, Guyana, Bolivia, and Ecuador. The dataset consists of 40,479 training images and 61,191 test images. The training set consists of the first 32,000 training images and our validation set consists of the remaining training images. [15]

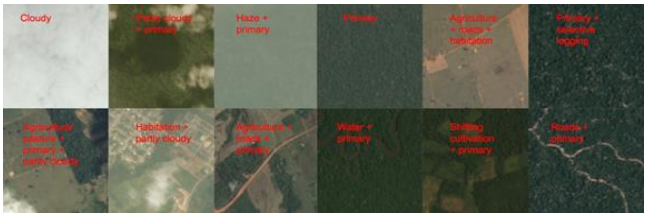


Figure 3. Sample satellite image data from Kaggle datasets with one or more label class.

The problem involves classifying satellite images as one or more of 17 labels. These labels are as follows:

Table 1. Classification of Images based on Geo-Atmospheric Difference

Atmospheric	Common Lands	Rare Land
Cloudy	Primary	Slash_and_burn
Partly_cloudy	Water	Selective_logging
Hazy	Habitation	Blooming
Clear	Agriculture	Conventional_mining
	Road	Artisanal_mining
	Cultivation	Blow_down
	Bare ground	

For the classification process, we propose the CNMSVM deep learning classifier. We compare the results with some of the basic classic classifier known such as support vector machine (SVM), modified support vector machine (mSVM), and a popular deep learning convolutional neural network (CNN).

4.2 Proposed Convolutional Neural modified Support Vector Machine

The proposed CNMSVM follows the following algorithm based on the enhanced modified SVM with feature space intercept (FSI), decision structure, and ideal boundary estimation.

CNMSVM Algorithm

1. Train an input image
2. Select the region of interest
3. Apply modified SVM to fine tune input data and group similar features of the images enhancing marginal neighborhood determination of support vectors
4. Apply initialization of all weights and biases of the CNN to a small value.
5. Integrate the fine-tuned data from modified SVM for CNN training.
6. Propagate pattern through the network through layers.
7. Back propagate if error factor is established.
8. Update weights and biases.
9. Find Mean Square Error (MSEI) until it reached maximum bounds.

4.3 SVM Modification: FSI and Balanced DS

Since in SVM, three parameters affect the decision outcome which is the b, α_i, and K it is only but ideal to modify K based on the training distribution and kernel-boundary-alignment offers to address data imbalance problem. Thus, we provided a model map featuring the innovative approach of modifying how SVM filters input elements to pre-tune datasets for better CNN input.

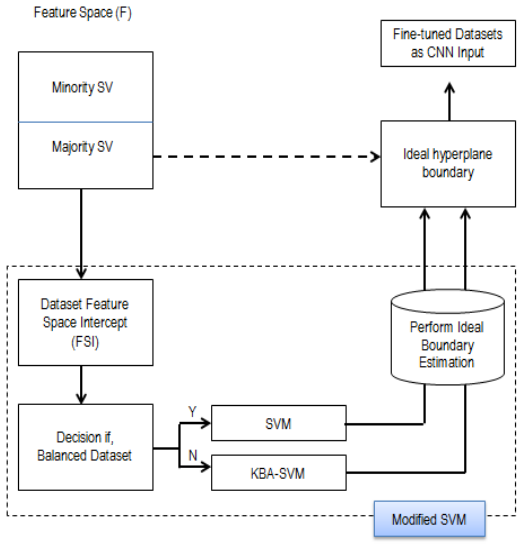


Figure 4. Map diagram of the proposed model with feature space intercept (FSI), decision structure, and ideal boundary estimation.

5. EXPERIMENTS AND RESULTS

Because of the complexity of the model, the training process takes longer time but could result in higher accuracy. Also, the input images of CNMSVM need to be the size of 224×224×3 to accommodate the input layer of CNMSVM.

In evaluating the results based on the given algorithms an important component was played by the F_2 score giving effectiveness of the classifier. Using precision p and recall r, defined as where p is the ratio of true positives (tp) to all predicted positives (tp + fp), r is the ratio of tp to all actual positives (tp + fn):

$$(1+\beta^2)\ pr/(\beta^2 p+r) \quad (1)$$

$$p = \text{tp/tp+fp} \quad (2)$$

$$r = \text{tp/tp+fn} \quad (3)$$

For the F_2 score, $\beta = 2$ was presented. Therefore, the mean F_2 score is considered to be the average of individual F_2 scores for each row in the dataset. Note that the F_2 score weighs recall higher than precision.

5.1 CNMSVM Results

The use of different classification model as baseline such as the support vector machine and convolutional neural network shows that the F_2 score improves through utilization of deep learning applications.

It can be noted that using F_2 score to evaluate each model is based on the premise that F_2 score is known to weigh recall higher than precision. Hence, in improving the said score there is a need to lessen the classification threshold after the sigmoid function. Tuning in the threshold value in the range of 0 to 0.5 in finding the optimal threshold value 0.2 this was used in all across the different classes.

As compared with SVM model, it can be directly noticed that there is an increase in the accuracy of F_2 score and such attribute was a significant result of CNN model. Further improvement can be seen as the proposed model was used obtaining better visualization of the target class labels.

Table 2. Image Classifications Scores on the Dataset using various Base Classifiers and CNMSVM

Model	F_2 Score
SVM	0.65631
mSVM	0.67346
CNN	0.81021
CNMSVM	0.91412

6. CONCLUSION AND FUTURE WORK

The use of satellite images based on multi-label classification problem posed several challenges especially in rendering large images and the accumulation of time understanding classification behavior and aiming for accuracy. This is a good learning as well as contributory on classifier standards where using the proposed CNMSVM architecture builds better label prediction although being complicated. The results are presented where high classification accuracy via F_2 score based on the deep learning approach provides a score value of 0.91412. Thus, the experimental results show that the best performance is obtained using the proposed CNMSVM architecture.

Further studies can be aimed at utilizing other satellite images either using the same algorithm or present with more sophisticated architectures such as ResNet and InceptionNet. Also, it is relevant to take advantage of unsupervised models such as SOM as a preprocessing algorithm or as an augmentation function for CNN

7. ACKNOWLEDGMENTS

Our deepest gratitude to the Commission on Higher Education and the Graduate Study Program of TIP.

8. REFERENCES

[1] Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification. IEEE Transactions on GeoScience and Remote Sensing, 55(2), 645–657.

[2] Chen, X., Xiang, S., Liu, C., & Pan, C. (2014). Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks. IEEE Geoscience and Remote Sensing Letters, 11(10), 1797–1801.

[3] Schlosser, J., Chow, C. K., & Kira, Z. (2016). Fusing LIDAR and Images for Pedestrian Detection using Convolutional Neural Networks. In IEEE International Conference on Robotics and Automation (ICRA) (pp. 2198–2205). Stockholm, Sweden.

[4] Sun, X., Xu, Z., Mengl, N., Lam, E. Y., & So, H. K. (2016). Data-driven light field depth estimation using deep convolutional neural networks. IEEE Journal, 367–374.

[5] Wolfshaar, J. van de, Karaaba, M. F., & Wiering, M. A. (2015). Deep Convolutional Neural Networks and Support Vector Machines for Gender Recognition. 2015 IEEE Symposium Series on Computational Intelligence, 188–195. https://doi.org/10.1109/SSCI.2015.37

[6] Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. IEEE Geoscience and Remote Sensing Letters, 14(5), 778–782.

[7] Wu, G. W. G., & Chang, E. Y. (2004). Aligning boundary in kernel space for learning imbalanced dataset. Fourth IEEE International Conference on Data Mining (ICDM'04). https://doi.org/10.1109/ICDM.2004.10106

[8] Nasr, A. H., & Helmy, A. K. (2009). Integration of Misrsat-1 and SPOT-2 Data for Quantitative Change Detection Applications. Graphics, Vision and Image Processing Journal, 9(5).

[9] Xu, E., & Zeng, O. (2017). Predicting Amazon Deforestation with Satellite Images. Stanford University. Retrieved from http://cs231n.stanford.edu/reports/2017/pdfs/916.pdf

[10] Ahmed, B., & Noman, A. Al. (2015). Land Cover Classification for Satellite Images based on Normalization Technique and Artificial Neural Network. In 1st International Conference on Computer & Information Engineering (pp. 26–27). Rajshahi, Bangladesh: Dept. of CSE, Rajshani University of Engineering & Technology.

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, 1097–1105. https://doi.org/10.1145/3065386

[12] Torres, J. (2016). First Contact with TensorFlow Part 1 : Basics.

[13] Miyazaki, H., Kuwata, K., Ohira, W., Guo, Z., Shao, X., Xu, Y., & Shibasaki, R. (2016). Development of an Automated System for Building Detection from High-Resolution Satellite Images. In Fourth International Workshop on Earth Observation and Remote Sensing Applications (pp. 8–11). IEEE Computer Society.

[14] Praveena, S., & Singh, S. P. (2015). Hybrid clustering algorithm and Neural Network classifier for satellite image classification. In International Conference on Industrial Instrumentation and Control (ICIC) (pp. 1378–1383). Pune, India: IEEE Computer Society.

[15] Planet: Understanding the amazon from space

Brain Tumor Segmentation on MR Images Using Anisotropic Deeply Supervised Convolutional Neural Network

Md Minhazul Islam
College of Information Science & Technology
Donghua University
Shanghai,China
minhazulislam08@gmail.com

Zhijie Wang
College of Information Science & Technology
Donghua University
Shanghai,China
wangzj@dhu.edu.cn

Muhammad Ather Iqbal
College of Information Science & Technology
Donghua University
Shanghai,China
m.atheriqbal@gmail.com

Guangxiao Song
College of Information Science & Technology
Donghua University
Shanghai,China
1169199@mail.dhu.edu.cn

ABSTRACT
Glioma is the most common and cancerous type of early stage brain tumors that arise from glial cells. Deep learning based solutions are flexible in brain tumor image analysis and computer-assisted diagnosis but it does not gain the desired accuracy. Manual brain tumor segmentation is a challenging task for clinicians therefore, researchers are working continuously to improve the accuracy for brain tumor segmentation using automatic segmentation. In this paper we segmented the brain tumor in three regions namely whole tumor, enhancing tumor and non-enhancing tumor using Convolutional Neural Network (CNN) implemented by anisotropic dilated convolution with residual blocks additionally using deeply supervised layers. Experiment shows that, our proposed method achieved the dice scores of 0.91, 0.78, 0.84 for whole tumor, enhancing tumor and non-enhancing tumor respectively, which is better than the BraTS 2017 challenge and other reported approaches.

CCS Concepts
- **Computing Methodologies ~ Image and Video Acquisition**

Keywords
Brain Tumor Segmentation; Anisotropic Convolution; Dilated Convolution; Deeply Supervised; Convolutional Neural Network

1. INTRODUCTION
A human brain can contain several types of tumors, among which one of the most common is glioma. Clinicians divide the gliomas into two types; namely, high-grade gliomas (HGG) which is considered to be more aggressive and life expectancy is an average of 2 years, on the contrary low-grade gliomas (LGG) is considered to be less aggressive and the life expectancy is up to several years.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301525

In medical imaging, Magnetic Resonance Imaging (MRI) offer bundle of image slices for diagnosis of brain tumor. The MRI image can be viewed by four sequences like T1-weighted, contrast enhanced T1-weighted (T1c), T2-weighted and Fluid Attenuation Inversion Recovery (FLAIR).

Brain tumor segmentation is a critical task, because of its complex architecture. Besides that, the tumors may form anywhere in the brain with different shape and size, while the MRI test scans also contrast poorly. Manual segmentation is tedious and very time consuming for clinicians and radiologists. On the contrary, automatic segmentation can make a great impact due to which many researchers tend towards automatic segmentation which is more convenient.

From the last couple of decades, many methods are used for the brain tumor image segmentation. Some approaches depend on the probabilistic atlases. The sample of the probabilistic distribution is Markov Random Field (MRF) [1]. MRF mostly segment and classify with the labels of brain images. MRF results in smoother segmentation that collects useful information from the neighborhood voxels. Due to some disadvantages several authors use classifiers in MRF like Support Vector Machine (SVM) [2], [3] and Random Forests (RF) [4], [5]. Marco et al. evaluated one-versus-one and one-versus-all classifiers based on SVM [6].

In recent years, deep learning is used as a novel method to improve the brain tumor segmentation accuracy with less time. In general, learning procedure uses neural network as the principle of deep learning. Convolutional Neural Network (CNN) is widely accepted in medical imaging. It can segment 2D slices without considering 3D circumstantial information. In the area of brain segmentation Zhang et al. and Kamnitsas et al. [7], [8] segmented tissue and brain tumor by CNN respectively. Havaei et al. [9] segmented the white matter lesion reducing the effect of omitted modalities. DeepMedic [10] uses 11 layers 3D CNN for brain tumor segmentation task. There are many other techniques used for brain tumor segmentation like 3D U-Net [11] which uses volumetric image segmentation for end-to-end training and testing. HighRes3DNet is based on end-to-end 3D CNN architecture using dilated convolution and residual blocks. Urban et al. [12] used 3D convolutional kernels for the segmentation of tumor. Zikic et al. [13] proposed a sliding-window concept in the 3D space for CNN. Pereira et al. [14] investigated brain tumor using CNN with small 3x3 kernels. Lyksborg et al. [15] suggested an ensemble of 2D CNN for tumor segmentation. Nowadays, Brain Tumor Segmentation (BraTS) challenge gained popularity in brain tumor

segmentation. Every year this challenge is getting famous due to great advancement and improvement in results.

In this paper, we propose a novel method to segment brain tumor using anisotropic dilated deeply supervised network [16]. In the beginning, we segment the effected tumor tissues step by step using a cascaded network to precise edema, non-enhancing tumor and enhancing tumor. The advantage of using cascaded architecture in tumor sub-region is to reduce false positive. In the next step anisotropic network improves 3D shapes affected by different alteration and improves the segmentation accuracy. The last and novel part of our proposed work is deeply supervised network, which learns features and vanishing gradient productively. It represents superior performance to control the training procedure as a result gives maximum similarity along with the ground truth and segmented result.

2. METHOD

Receptive field or field of view is one of the most important parts of deep CNNs - how much a basic image can see a neuron in a certain layer? It depends on receptive field. A larger receptive field may create a problem for larger memory consumption, model complexity, image resolution and number of features in the network. Due to these reasons we are using anisotropic convolutional neural network to get large receptive field as input and small receptive field in the output. In our work the receptive fields for whole tumor, tumor core, and enhancing tumor are 217×217, 217×217 and 113×113 respectively. We decompose a 3D filter with a size of 3×3×3 into an intra-slice filter with a size of 3×3×1 and an inter-slice filter with a size of 1×1×3, in order to deal with the anisotropic receptive fields.

Our proposed network architecture is shown in Figure 1. We use 10 residual blocks with anisotropic convolution, dilated convolution and specially added deeply supervised network. Whole tumor, tumor core and enhancing tumor uses twenty intra-slice and four inter-slice convolutional layers. It also uses two down sampling layer for whole tumor and tumor core because of its larger size from enhancing tumor. Considering its smaller input size enhancing tumor uses one down sampling layer. Every residual block has two intra-slice convolutional layers and each block is connected directly to the output for learning residual function. It is very complicated to know the actual depth of deep network. Layers can be too deep or too narrow, for the former case it is difficult to correct the error and for the latter case it is difficult to have enough representation power. Though, in deep residual network [17], without worrying about the degradation problem we train very deep layers safely which leads to get enough learning power. To enlarge the receptive field we used dilated convolution. The red boxes show the residual blocks with dilation, brown block represent the downsampling layer and the blue box shows the deeply supervised layer in Figure 1.

In the field of brain image segmentation task, we can get more accurate segmentation results by implementing deep CNN having multi-scale steps with different strides. From the earlier step to the later steps, the hidden layers would transfer the multi-level and the multi-scale features. The extracted features at hidden layers are less meaningful as they lack deep supervision. Moreover, some of the meaningful information was not fully utilized and lost.

To overcome such problems, we propose to add deeply supervised layers [16], which implement straight and early regulation for both hidden layers and output layer. This way we can efficiently reduce prediction error and minimize classification error. We use three deeply supervised layer for whole tumor and tumor core and two deeply supervised layer for enhancing tumor. These five deeply supervised layers consist of one channel convolutional layer with filter size 1×1. It controls both learning features and the gradients of back propagation while training. It compares the ground truth and segmentation result.

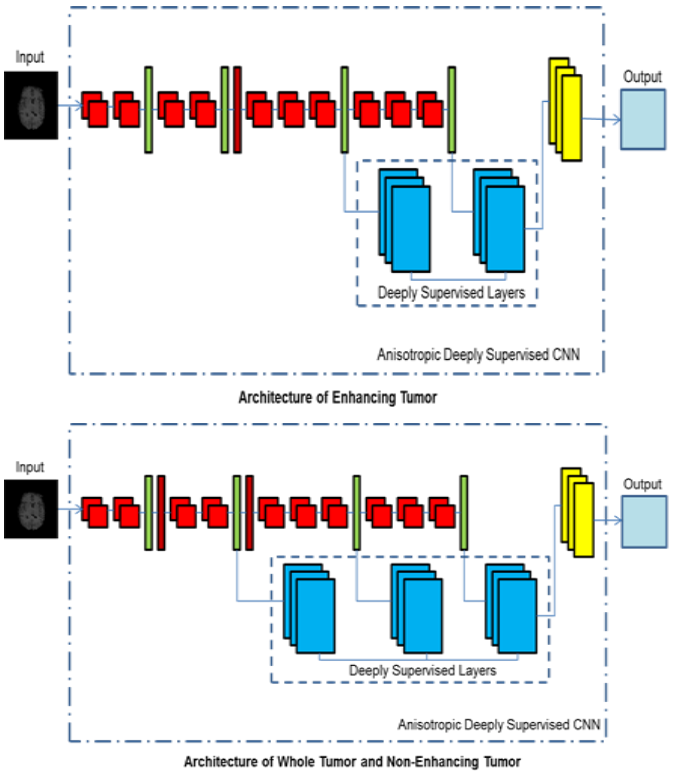
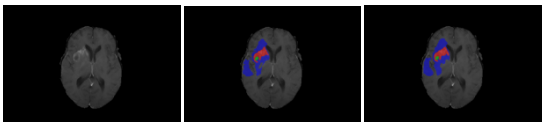
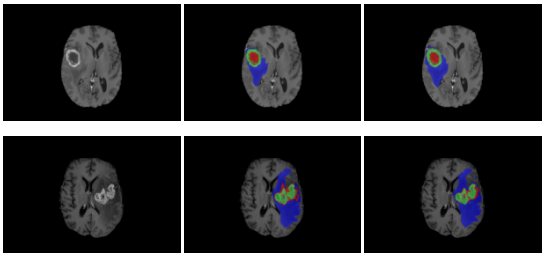


Figure 1. Our proposed architecture of anisotropic deeply supervised network with dilated convolution and residual connection. Because of its smaller input size enhancing tumor use only one downsampling layer.

3. EXPERIMENTAL SETUP

In a couple of years, every year a challenge held about brain tumor segmentation namely BraTS as a part of MICCAI conference. In our paper we have used BraTS 2015 dataset [18]. The images were taken from the real patients. It contains about 300 high- and low-grade glioma cases. The voxel resolution of each image slice is $1mm^3$. Each image scan contains 4 modalities, namely T1, T1c, T2 and FLAIR. The MRI images have four parts as like necrosis, edema, enhancing tumor and non-enhancing tumor. The dataset was divided into two parts; training dataset including ground truth and testing dataset. The training dataset contains 220 HGG patients and 54 LGG patients. These were mixed with 53 high-grade and low grade gliomas for testing. The testing dataset is segmented manually by human experts (one to four raters). The annotations were visually inspected and approved by qualified raters. We execute our network in TensorFlow [19] using NiftyNet [20]. The initial learning rate is set to 10^{-3}, weight decay 10^{-7}, batch size 2 and maximum iteration 20k.

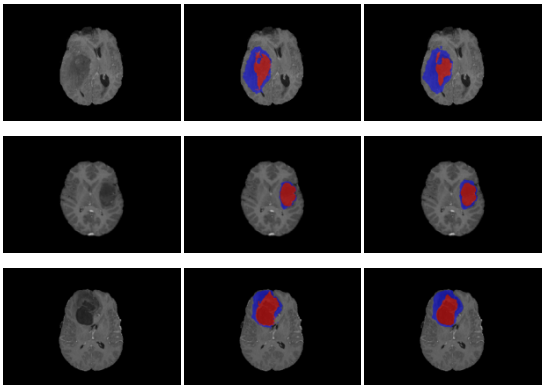




T1 Image Ground Truth Image Segmented Image

Figure 2. Visual segmented result from our proposed anisotropic deeply supervised CNN model that shows the T1 MR images of high-grade gliomas (HGG) with the ground truth. Blue, green and red colors represent the whole tumor, enhancing tumor and non-enhancing tumor respectively.

NVIDIA GTX 1070 GPU is used for training. The training patch of whole tumor was 144×144×19; tumor core was 96×96×19 and for enhancing tumor 64×64×19. Each network was trained by using Dice loss function. For visualization and understanding we used ITK-SNAP-3 software.



T1 Image Ground Truth Image Segmented Image

Figure 3. Visual segmented result from our proposed anisotropic deeply supervised CNN model that shows the T1 MR images of low-grade gliomas (LGG) with the ground truth. Blue, green and red colors represent the whole tumor, enhancing tumor and non-enhancing tumor respectively.

4. RESULTS AND DISCUSSION

For evaluating the lesion segmentation models, we report the testing results in terms of "Dice" (a widely used metric). Dice score can be calculated as:

$$Dice = \frac{2\mathrm{TP}}{\mathrm{FN} + \mathrm{FP} + 2\mathrm{TP}} \qquad (1)$$

where TP, FN and FP stands for "True positive", "False negative" and "False positive" predictions, respectively. It helps to determine the overlap area of the predicted lesion region and the ground truth.

We show that T1c image slices for both high-grade glioma and low-grade glioma includes whole, enhancing and non-enhancing tumors, respectively, In Figure 1 and Figure 2, the blue, red and green colors show the areas of whole tumor or edema, enhancing tumor and non-enhancing or tumor core, respectively. We compared our dice score with BraTS 2017 and our dice score is better than the highest dice score of this challenge. The evaluation of dice score is shown in Table 1.It shows the mean, median and standard deviation values. Our anisotropic deeply supervised CNN gives the dice scores of 0.91, 0.79 and 0.84 for whole, enhancing and tumor core, respectively. Mean Dice score of other baseline approaches are shown in Table 2. We constructed our architecture with anisotropic and dilated convolution with residual connections added with deeply supervised layers. During training the hidden layers gives better performance as we use deeply supervised layers. The back propagation process in these deeply supervised layers accelerates the parameters update but also avoiding the important information lost. Our work shows that it is very efficient to achieve the state of the art result in the field of brain tumor segmentation.

Table 1. Dice score of our proposed work on BraTS 2015 dataset.

	Dice Score		
	Whole Tumor	**Non-Enhancing Tumor**	**Enhancing Tumor**
Mean	0.91	0.78	0.84
Standard Deviation	0.113	0.192	0.15
Median	0.93	0.89	0.85

5. CONCLUSION

In this paper, the proposed model with anisotropic dilated convolution, residual connection and additionally deeply supervised layers segments the MRI brain image slices more accurately as obtained our dice scores showed in Table 1. Anisotropic convolutional layers work with the receptive field to get competitive performance and balance among the model complexity, memory consumption and receptive field. Dilated convolution is responsible to enlarge the receptive field. The residual connection speeds up the convergence of training. The additional deeply supervised layers play a key role in our model. It is used to control the difference between the segmentation result and ground truth. This makes our trained architecture robust. The learning process of hidden layers minimizes the classification error by deeply supervised layers. Experiment shows that, our proposed method achieved the dice scores of 0.91, 0.78, 0.84 for whole tumor, enhancing tumor and non-enhancing tumor respectively which is better than the Brats2017 challenge and other reported approaches.

Table 2. Comparison of mean Dice scores of different baseline approaches using BraTS 2013, 2014, 2015 and 2017 dataset.

S. No	Publication	Dice Score		
		Whole Tumor	Non-Enhancing Tumor	Enhancing Tumor
1	Urban et al. [12]	0.87	0.77	0.73
2	Zikic et al. [13]	0.84	0.74	0.69

3	Davy et al. [21]	0.85	0.74	0.68
4	Dvork and Menz [22]	0.83	0.75	0.77
5	Havaei et al. [9]	0.88	0.79	0.73
6	Lyksborg et al. [15]	0.80	0.64	0.59
7	Kamnitsas et al. [8]	0.85	0.67	0.63
8	Wang et al. [23]	0.90	0.83	0.78

6. REFERENCES

[1] Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.A., Ayache, N. and Golland, P.: A generative model for brain tumor segmentation in multi-modal images. In: MICCAI pp. 151-159 (2010)

[2] Bauer, S., Nolte, L.P. and Reyes, M.: Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization. In: MICCAI pp. 354-361 (2011)

[3] Lee, C.H., Wang, S., Murtha, A., Brown, M.R. and Greiner, R.: Segmenting brain tumors using pseudo–conditional random fields. In: MICCAI pp. 359-366 (2008)

[4] Geremia, E., Menze, B. and Ayache, N.: Spatially Adaptive Random Forest. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro pp. 1332-35 (2013)

[5] Pinto, A., Pereira, S., Correia, H., Oliveira, J., Rasteiro, D.M. and Silva, C.A.: Brain tumour segmentation based on extremely randomized forest with high-level features. In: EMBC pp. 3037-3040 (2015)

[6] Marco, S. B., Carlos, C. G., Juan, S. B., & Kathleen, S. V. Brain Tissue Model Classification for Telesurgery Navigation. International Journal of Machine Learning and Computing, 5(1), 68 (2015)

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770-778 (2016)

[8] Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36, 61-78 (2017)

[9] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural net- works. Medical Image Analysis 35, 18-31 (2016)

[10] Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36, 61-78 (2017)

[11] Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI, pp. 424–432 (2016)

[12] G. Urban, M. Bendszus, F. Hamprecht, and J. Kleesiek, "Multimodal brain tumor segmentation using deep convolutional neural networks," in MICCAI BraTS (Brain Tumor Segmentation) Challenge. Proceedings, Winning Contribution, pp. 31–35, Boston, MA, USA, (2014)

[13] D. Zikic, Y. Ioannou, M. Brown, and A. Criminisi, "Segmentation of brain tumor tissues with convolutional neural networks," in Proceedings MICCAI-BRATS, pp. 36–39, Boston, MA, USA, (2014)

[14] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1240–1251 (2016)

[15] M. Lyksborg, O. Puonti, M. Agn, and R. Larsen, "An ensemble of 2D convolutional neural networks for tumor segmentation," in Scandinavian Conference on Image Analysis. SCIA 2015, vol 9127, Lecture Notes in Computer Science, Springer, Cham, (2015)

[16] C. Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply supervised nets. Eprint Arxiv, pages 562-570 (2014)

[17] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)

[18] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, C¸ ., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K.: The multimodal brain tumor image segmentation benchmark (BRATS). TMI 34(10), 1993–2024 (2015).

[19] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., Brain, G.: TensorFlow: A system for large-scale machine learning. In: OSDI, pp. 265–284 (2016)

[20] Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Barratt, D.C., Ourselin, S., Cardoso, M.J., Vercauteren, T.: NiftyNet: A deep-learning platform for medical imaging (2017). arXiv preprint arXiv:1709.03485 (2017)

[21] A. Davy, M. Havaei, D. Warder-Farley et al., "Brain tumor segmentation with deep neural networks," in Proceedings MICCAI-BRATS, Boston, MA, USA, (2014)

[22] P. Dvorak and B. Menze, “Structured prediction with convolutional neural networks for multimodal brain tumor segmentation,” in Proceedings MICCAI-BRATS, pp. 13–24, Munich, Germany, (2015)

[23] Wang, G., Li, W., Ourselin, S., & Vercauteren, T. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In International MICCAI Brainlesion Workshop pp. 178-190, Springer, Cham. (2017)

Automatic Prediction of the Conversion of Clinically Isolated Syndrome to Multiple Sclerosis Using Deep Learning

H. M. Rehan Afzal
University of Newcastle, Hunter Medical Research Institute
University Drive
Callaghan NSW Australia
c3249813@uon.edu.au

Suhuai Luo
University of Newcastle
University Drive
Callaghan NSW Australia
suhuai.luo@newcastle.edu.au

Saadallah Ramadan
University of Newcastle, Hunter Medical Research Institute
University Drive
Callaghan NSW Australia
saadallah.ramadan@newcastle.edu.au

Jeannette Lechner-Scott
University of Newcastle, Hunter Medical Research Institute, Department of Neurology, John Hunter Hospital
University Drive
Callaghan NSW Australia
jeannette.lechnerscott@newcastle.edu.au

Jiaming Li
CSIRO Data61,
Marsfield, NSW 2122, Australia
Jiaming.li@data61.csiro.au

ABSTRACT

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous. Disability can be prevented by early detection of lesions. Deep learning techniques, such as convolutional neural networks (CNN), can learn patterns on brain magnetic resonance image (MRI) so as to predict the conversion of clinically isolated syndrome (CIS) to definite multiple sclerosis. The aim of this paper is to develop a method to automatically detect the conversion. The proposed algorithm is an improved CNN which uses LeNet architecture coding in Python and Keras library. It consists of different convolutional layers which learn the patterns of input images by using convolutional filters. ReLU activation function and max-pooling are used to reduce the dimensions of images for efficient and fast processing. A detailed investigation of automatic prediction algorithm is performed on the MRI images of 21 patients. The 21 patients were scanned at onset of CIS and a year later (of whom, 11 converted to MS and 10 did not convert to MS). The proposed deep learning algorithm predicted the presence of MS with an accuracy of 83.3% and 100% in two experiments. In the first experiment, 5 out of 6 patients were predicted correctly. In the second experiment, 6 out of 6 patients were predicted correctly. The experiments have proved that the proposed method is an automated system which can predict the disease accurately and quickly, contributing to the prevention and alleviation of the disability caused by the disease.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301526

CCS Concepts

• Computing methodologies~Vision for robotics

Keywords

Multiple sclerosis, Deep learning, CNN, MS Lesions, CIS, Prediction of MS.

1. INTRODUCTION

Multiple sclerosis (MS) is considered as a chronic disease of the central nervous system (CNS) which results into inflammation and demyelination of nerves. Basically, it disrupts the communication between the body and brain. According to world health organization (WHO), more than 2,500,000 patients have multiple sclerosis worldwide. During 2009, it was estimated by MS Australia that 25,000 patients were registered in Australia alone and its prevalence is increasing rapidly [1]. There is ample evidence that if MS is detected early, the long-term prognosis can be improved [2].

The first episode of MS is named clinically isolated syndrome (CIS) as the diagnosis of MS requires dissemination in time and space [3]. The majority of CIS will progress to multiple sclerosis but after what time interval differs from case to case. The sooner the case progresses to clinically definite MS, the worse the prognosis. So timely prediction of conversion to clinically-definite MS (CDMS) and early treatment intervention can save future damage to the central nervous system and ultimately prevent disability. According to Fisniku et al. [3], about 30% of MS patients with CIS will get second clinical attack within one year, indicating CDMS.

The accurate identification of patients suitable for aggressive disease modifying treatment is considered very difficult and critical. MRI is the standard diagnostic tool and plays an important role in monitoring progression and prognosis of the disease. MRI protocols for MS detection are proposed by various expert panels. For this purpose, different criteria are made by

different organizations. Among these criteria, the McDonald [4] and MAGNIMS [4] guidelines are most widely accepted.

The MAGNIMS guidelines and McDonald's criteria help us in the prediction of CDMS from CIS and long-term disability [4]. For the assurance of highest specificity and sensitivity, McDonald has set guidelines which are referred as McDonald's criteria. It provides instructions to diagnose MS. Lesion detection is best done on T2 FLAIR, whereas gadolinium administration can help in differentiation between acute and chronic lesions. T2-weighted MR images are considered the most sensitive analytic test for indicating the dissemination of disease. So, by keeping all these guidelines in mind, we proposed an automated algorithm to predict the conversion of CDMS from CIS.

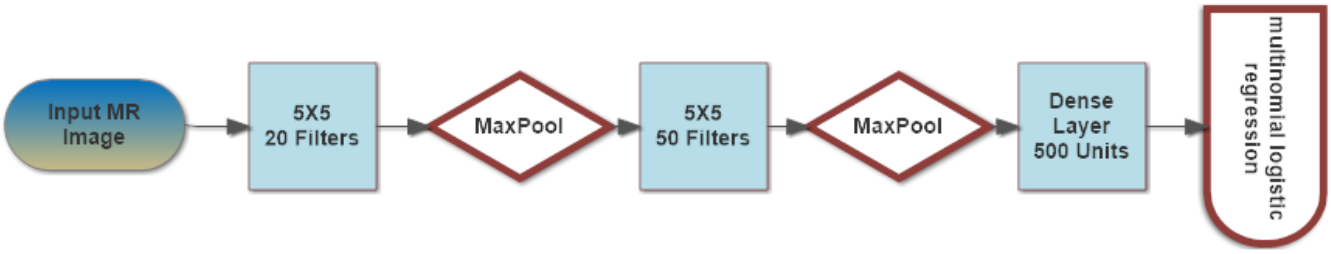


Figure. 1. The Proposed Algorithm

Deep learning techniques with CNN are used to design such algorithm which can learn patterns and features of lesions by multinomial logistic regression to predict MS. The proposed algorithm can be seen in figure 1.

The proposed algorithm will be able to efficiently and automatically predict the conversion within one-year, telling which patient will convert to CDMS and which will not be leading to more rapid implementation of appropriate therapy. The rest of the paper is summarized as follows. Section 2 describes the proposed method. Results are evaluated in section 3, and conclusion is given in section 4.

2. The Proposed Method

The proposed method is implemented using Keras Library [5] with deep learning in Python. It mainly uses LeNet architecture [6]. The main flow of algorithm is shown in figure 1. Images used as input are scanned with Siemens MAGNETOM Prisma 3T unit [7]. The images were taken from John Hunter Hospital's (JHH) dataset. JHH ran a clinical trial which scanned images at baseline (CIS) and then after one year. It gathered T2-w images of 21 patients, with baseline and after one year of follow-up. Manual assessments are performed by qualified neurologists.

By using these images, we implemented a deep learning prediction model which performs prediction with the help of conventional neural networks. The proposed method consists of data augmentation, CNN architecture, training architecture and testing architecture, as described below.

2.1 Data Augmentation

Data augmentation technique is used when we have limited data, mostly while dealing with medical images. Medical data is normally of less quantity, which is a difficult part to handle because deep learning always needs large volumes of data. This technique adds value to already existing data by performing different tasks like flipping, cropping, rotating, translating and scaling, etc. It helps deep learning techniques for training purposes.

While training the deep learning model, we are tuning its parameters, so it could map input to a specific output. The main goal is to make the model loss the least possible, which is only possible when these parameters are tuned in the right way. So, it is clear that if we have more parameters, our training algorithm would have more examples to learn and the performance would be good. Therefore, to overcome the shortage of data, we have applied augmentation technique which consists of flipping, cropping, rotating, translating and scaling.

2.2 CNN Architecture

CNNs are being used in computer vision field for decades [8-10]. CNNs are becoming a popular deep learning technique now, not only for computer vision techniques but also for other numerous applications like hyperspectral imaging, language processing, detection and medical image processing. CNNs for medical imaging are being used since 1990s. Several examples [11-14] of medical image analysis are detection of lung cancer, Alzheimer's disease detection and automated systems for detection of microcalcification, etc. Because of the availability of high-performance GPUs and intense research, computer aided diagnosis's have shown excellent performance which not only save time for doctors but also provide efficient future prediction of diseases. CNNs consist of convolutional layers. These layers perform local features detection for the input image in all locations. Every layer ends up with a set of nodes which are further connected with a subset of neurons. So, for the detection of local features which have same properties, different weighs are shared between the convolutional layers.

2.3 The Proposed CNN Architecture

In the proposed algorithm 20 filters of size 5 by 5 are used, which slide over the whole image and take a dot product. These filters are randomly initialized. As these nodes are in the hundreds and thousands, it makes our system very complex. To overcome the complexity, we have used pooling. Normally, max pooling is used for the reduction of features. The proposed architecture is a hierarchal procedure which can be shown in figure 1 which consists of convolutional layers, max pooling, a fully connected layer and some probability techniques at the end like logistic regression. Logistic regression technique is also called softmax layer, is a part of convolutional neural network, which helps us to generate desired output.

The CNN architecture used in the proposed method consists of 6-layer. It has two convolutional layers and two maxpooling. The first convolutional layer has 20 filters with the kernel size of 5X5. The second convolutional layer has 50 filters with the same kernel size of 5X5. Here, a relatively small number of convolutional filters are utilized to overcome the issue of overfitting. ReLu activation function is used to speed up the process. Then it is

followed by fully connected dense layer of LeNet architecture. At the end, multinomial logistic regression is used which performs probability task. It will return a list of probabilities by calculating the weights of each class and higher probability will be considered as final classification.

2.4 Training Architecture

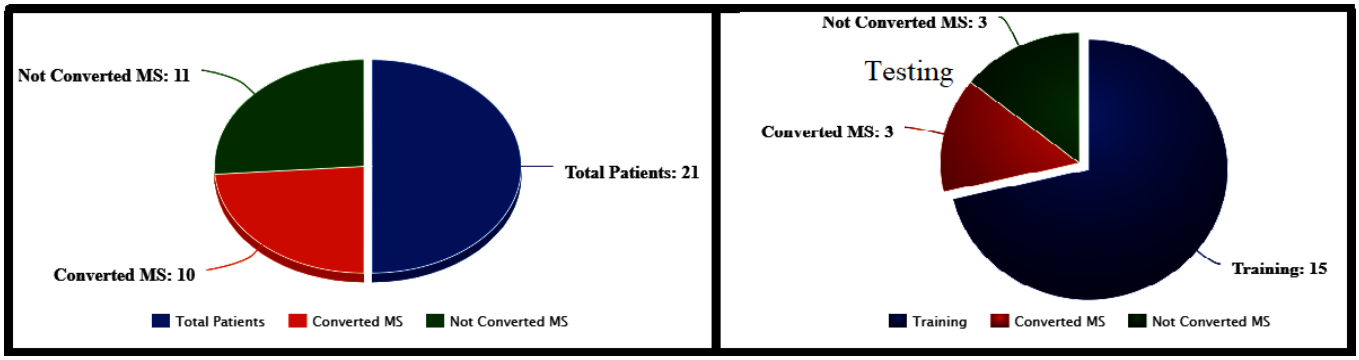
According to existing methods [15-17], it is clear that training depends on the available data. If large amount of data is available, then more CNN layers are needed which improves the accuracy during the classification process. To increase the number of layers means that parameters are also increased. But in medical field or specifically in multiple sclerosis field, the data available is very limited. MRI images have slices, ranging from 10 to 100 per patient depending on the scan. Therefore, in the field of medical imaging, we may need to add some techniques to overcome this issue. We applied data augmentation technique to increase the number of images for better training of our deep learning algorithm.

Data augmentation is considered an effective method which helps us increasing the precision of CNNs [18]. After data augmentation, our data set has been increased but not perfect as it needed to be for training purposes. As we discussed above, these limitations can affect the architecture and results in overfitting due to limited data. To handle this issue, we used fewer layers and less parameters which are discussed in the evaluations section. One thing was kept in mind while training with less parameters that accuracy of learning features and prediction must not be compromised.

2.5 Testing Architecture

The final step consists of a testing stage. After the training is completed, unseen images are used for testing purposes to check the accuracy of the proposed method. Testing was done two times by randomly chosen images. As we are limited to 21 patients, we have first taken randomly selected 15 patients for training and 6 patients for testing stage (3 converted to CDMS and 3 were not converted to CDMS). For the second go, we selected the training images from another 15 different randomly selected patients and the rest of the 6 patients were used for testing purpose. The purpose of these two experiments was to check the accuracy of the proposed method. All these patients were assessed by qualified neurologists at the JHH.

Details of used data can be seen in figure 2 (a) and (b).



a) Total Patients under the study of JHH b) Training and testing of Patient's data

Figure. 2. Pie Chart representation of Patient's Data used

3. Evaluations

The data used in the proposed method is taken from JHH, the MS clinic. Scans are usually performed at baseline 6m and 1 year after onset. We gathered 21 T2-w images with baseline and after one year of follow-up. These assessments were performed by qualified neurologists and radiologists. All patients included fulfilled the McDonald's criteria. Out of these 21 patients, 10 converted to CDMS after one year, whereas 11 did not convert to CDMS after one year follow up. At input, the baseline images (CIS) and one year of follow-up are used together to train the algorithm. Experiments were set in such a way that two experiments were performed for testing and training stage. At first go, images of 15 patients were used for training and 6 were tested. Later with one more go, different 15 patients were used for training stage whereas rest of 6 patients were used at testing stage to check accuracy of proposed algorithm.

3.1 Implementation

The proposed automated algorithm is implemented in Python. All training and testing experiments were run on Intel Core 5, 7th generation with 16 Gb RAM memory. LeNet architecture was used with ReLu activation function. 50 epochs were simulated at training stage.

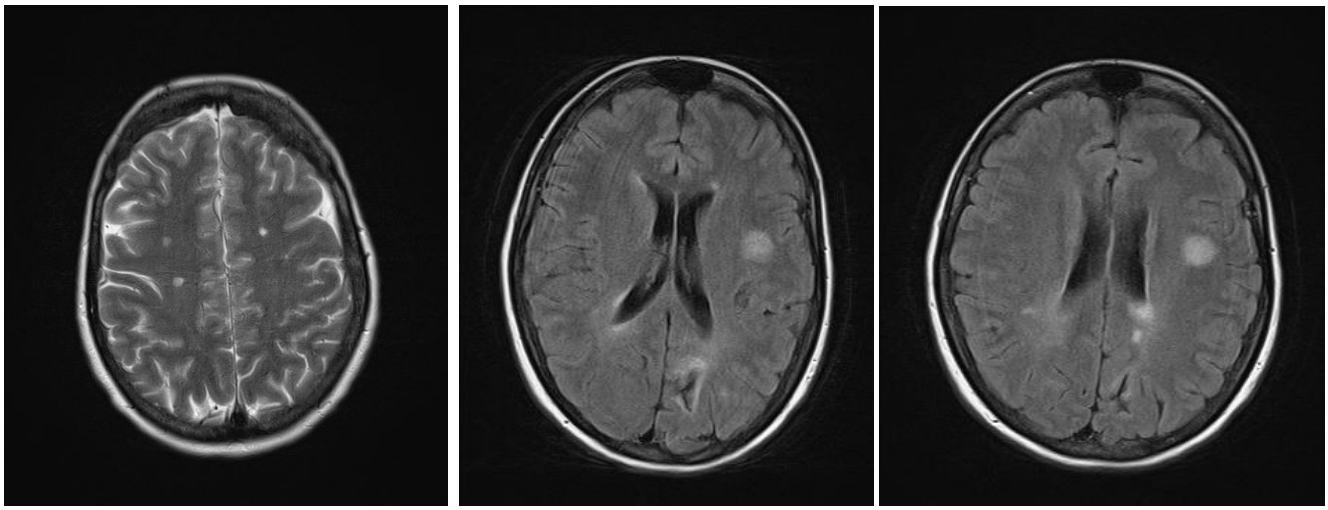
3.2 Results

Before running the algorithm, all images were assessed manually by qualified radiologists to confirm the exactness of algorithm. After running this automated algorithm, comparison was done with the manually verified images of patients as converted and non-converted to CDMS. The first experiment gives us 5 correct results out of 6 under testing. Some images of these patients can be seen in figure 3.

By inspecting the above 6 patients according to manual assessment, figure 3 (a) shows the images of those patients who converted to CDMS within one year and figure 3(b) shows the images of those patients who were not converted to CDMS after one year. When we run our automated algorithm with first go on these images, 5 patients were correctly classified

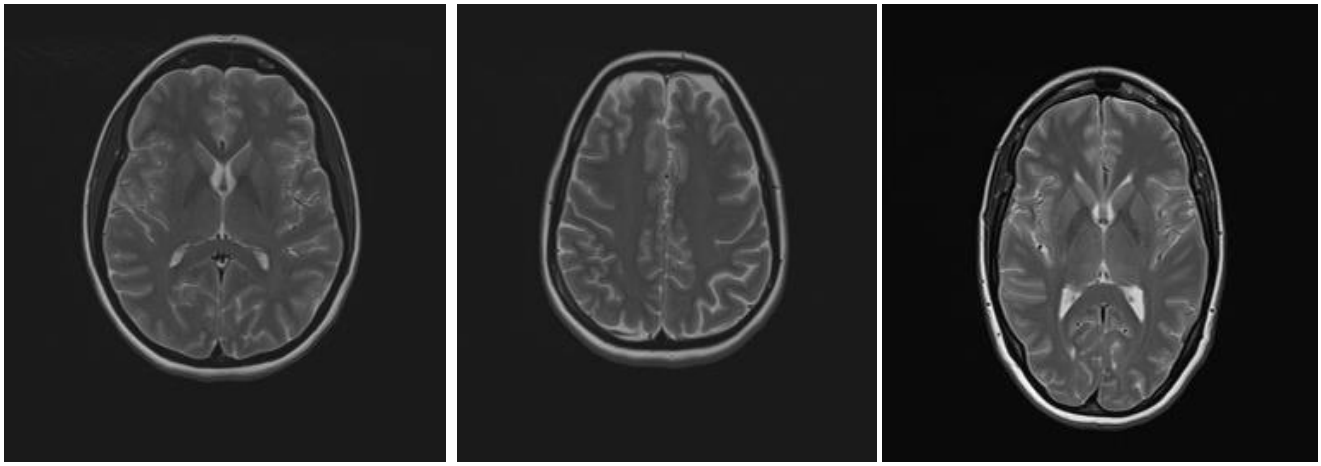
Table 1. Manual and automated classified results

Total patients: 21 (10 Converted & 11 Not-Converted)		
Experiment	15 training, 6 testing	Accuracy
First Go	5 correct out of 6	83.3
Second Go	6 correct out of 6	100



Patient 1 Patient 2 Patient 3

a**) Converted to CDMS within one year**



Patient 4 Patient 5 Patient 6

b) Not converted to CDMS after one year

Figure. 3. Results of proposed method

If we put deep insight on these images, the first patient in the figure (a) has converted to CDMS according to manual classification but our algorithm didn't classify correctly, whereas other 5 images were classified correctly, which gives 83.3% accuracy. Due to limited images, we randomly selected different 15 images for training and 6 images for testing in the second experiment. and in this go we got 100% results, meaning 6 correctly classified out of 6. The complete results with comparison can be seen in table 1.

4. CONCLUSION

Prediction of disease is a very active research field in medical sciences. Many methods have been proposed with the help of computer vision. Multiple sclerosis is considered as a disease, for which rapid progress in research has occurred leading to ever more therapeutic options. According to WHO, more than 2,500,000 patients have multiple sclerosis worldwide and its incidence and prevalence is increasing every year. So, there is a desperate need to apply automated MS detecting algorithms which

could help doctors and radiologists to detect MS early and implement therapy more rapidly.

To contribute to the MS detection, an automated algorithm is proposed using deep learning. It can automatically predict which patients will convert to CDMS or not within the following year. The proposed algorithm is tested on 21 patients, scanned on MRI and assessed by qualified radiologists and a panel of doctors. We tested this algorithm in two experiments. First experiment gave 83.3% accuracy (5 correctly classified out of 6) and second experiment gave 100% accuracy (6 correctly classified out of 6).

These results show that the proposed algorithm can predict the conversion from CIS to CDMS within one year of follow up. It also shows that the proposed algorithm is robust in nature and computationally not very complex. In the future, more data will be collected, and more deep learning structures will be investigated to improve the performance.

5. REFERENCES

[1] https://www.msaustralia.org.au/

[2] L. K. Fisniku, P. A. Brex, , D. R. Altmann, K. A. Miszkiel, C. E. Benton, R. Lanyon, ... & D. H. Miller, "Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. Brain:, 131(3), pp. 808-817, 2008.

[3] X. Montalban, "CIS diagnostics and predictors of conversion to CDMS," Multiple sclerosis and related disorders, 3(6), 764, 2014.

[4] C. H. Polman, S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi,, ... & F. D. Lublin. "Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Annals of neurology," 69(2), pp. 292-302, 2011.

[5] M. Filippi, M. A. Rocca, O. Ciccarelli, N. De Stefano, N. Evangelou, L. Kappos, ... & C. Gasperini, "MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. The Lancet Neurology," 15(3), pp. 292-303, 2016.

[6] https://keras.io/

[7] http://deeplearning.net/tutorial/lenet.html

[8] https://hmri.org.au/news-article/scanner-installation-milestone-mri-centre

[9] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biol. Cybern., vol. 36, no. 4, pp. 193–202, 1980.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[12] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "A comprehensive computer- aided polyp detection system for colonoscopy videos," in Information Processing in Medical Imaging. New York: Springer, pp. 327–338, 2015.

[13] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in Proc. IEEE 12th Int. Symp. o Biomed. Imag., pp. 79, 2015.

[14] N. Tajbakhsh and J. Liang, "Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks," in Proc. MICCAI, 2015.

[15] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in Proc. MICCAI, pp. 411–418, 2013.

[16] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," "ImageNet Challenge," pp. 1–10, 2014.

[17] K. He, X. Zhang, S. Ren , and J. Sun, "Deep residual learning for image recognition," "arXiv preprint arXiv":1512.03385, 2015.

[18] M. Havaei,, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with Deep Neural Networks. Medical Image Analysis," 35: pp. 18–31, 2017.

[19] A. Z. Spector, Achieving application requirements. In Distributed Systems, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738, 1989.

Similarity Detection of Color Image based on Main Color Table

Wenjia Ding
Shenzhen Research Institute,
Wuhan University, China
+86 755-26656730
greenflowerwh@qq.com

Yi Xie
Shenzhen Research Institute,
Wuhan University, China
+86 755-26656730
11794@qq.com

Yulin Wang
Shenzhen Research Institute,
Wuhan University, China
+86 755-26656730
wangyulinwh@qq.com

ABSTRACT
With the wider use of huge amount of images, more reasonable retrieval algorithms are becoming more and more important. This paper proposed an image similarity detection algorithm for image content based retrieval. The algorithm can extract the main color of the image, and improve the traditional methods in similarity detection performance. Because the eigenvalues of different regions of the image may be quite different, only using the global features of the image as matching eigenvalues will lead to a large error in similarity judgment, so image segmentation technology has always been the key technology in color image processing. In this paper, a more advanced block technology is used to segment the image of different regions, which can effectively reflect the spatial information of different images.

CCS Concepts
•Security and privacy → Software and application security → Social network security and privacy.

Keywords
Color image processing; ISODATA clustering; Similarity detection.

1. INTRODUCTION
Digital image processing is widely used in many aspects. How to quickly and accurately retrieve the required information from the vast amount of images has become a hot issue [1]. Image retrieval can be divided into text-based image detection and content-based image retrieval according to the description content.

Text-Based Image Retrieval (TBIR) used keywords to describe the content of an image [2] [3]. These keywords or called free text sometimes need to be manually labeled, so large-scale image retrieval requires a lot of manpower and material resources. In addition, manual descriptions do not accurately represent the full content of an image in many cases, so they do not match well with certain types of images.

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong.

ACM ISBN 978-1-4503-6613-7/18/12…$15.00.
https://doi.org/10.1145/3301506.3306497

To overcome these problems, CBIR (Content-Based Image Retrieval) came into being [4] [5] [6]. Different from content-based image retrieval, CBIR does not require manual annotation of image content, but automatically extracts the visual content of each image as its index by computer. In content-based image retrieval, the extraction of eigenvalues and the calculation of similarity have always been the focus and difficulty of related technologies. In this paper, we studies what features are extracted from images, how to extract these features, and better similarity measurement. A method of feature extraction and corresponding image segmentation based on dominant color is proposed.

2. COLAR IMAGE SIMILARITY DETECTION BASED ON MAIN COLOR TABLE
Firstly, the features of images are extracted and stored as a specific feature vector in the database. When the user needs to retrieve the image, the same feature extraction method is used to extract the features of the target image, and to match them with the those in the database, as shown in Fig. 1.

Image segmentation can effectively reflect the information of each part of the image. Based on these segmented image blocks, the local and spatial features of the image can be reflected to make up for the gap between human and computer image recognition.

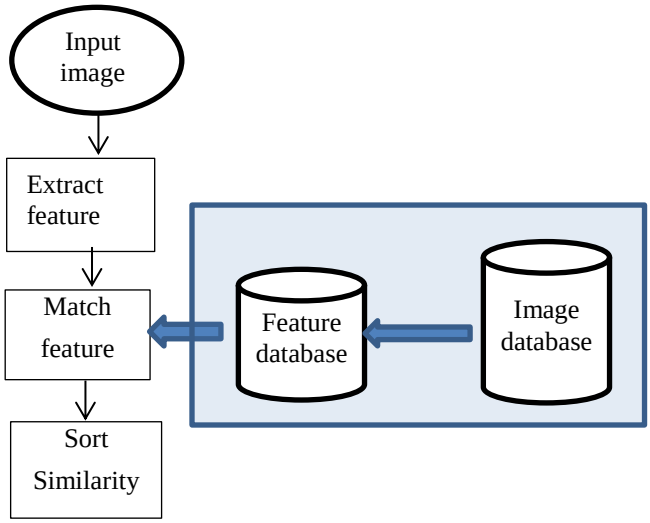


Figure 1. Content based image retrieval process

2.1 HSV Main Colors

Using image segmentation technology to divide an image into different locations of the region, and then to count the histogram of each region, to a certain extent, can solve the shortcomings of color histogram. The so-called main color is the highest proportion of colors in the image. It may be several to dozens of different colors. It is formed by clustering, so it is dynamic. Compared with RGB space, HSV space can express color sensitivity, tone and brightness more intuitively, which is more suitable for color contrast, and also convenient for emotional expression [7].

First, the R, G and B in the RGB model are transformed into H, S and V in the HSV model. Then, in order to be more consistent with the accuracy of the model, we do some preprocessing:

1) The color whose H, S and V are all 0 is set as black. At the same time, treat all colors whose V<15% as black, therefore set their H, S, V as 0.

2) The color whose H, S are all 0 and V is 1, is set as white. At the same time, treat all colors whose S<10%, and V>80% as white, therefore set their H=0, S=0, and V=1.

3) The others in between are color, and the values of H, S and V remain unchanged.

By treating the color similar to white as white, and the color similar to black as black, the accuracy of the model is improved.

2.2 ISODATA Clustering

We select the clustering method of ISODATA [8], and improve it to cluster the image. In the 24-bit true colors, R, G, and B all have 256 different colors. To facilitate calculation, we divide the 256 color values of these three colors into 16 color ranges on average, and calculate the number of pixels in different color ranges in the image, and arrange them from large to small. Finally, we select 100 colors of the largest number of pixels, then select Nc initial clustering centers from these 100 colors. The specific selection rules of clustering center are as follows:

Firstly, the color with the most pixels is taken as the first clustering center, then the color with the second most pixels is taken as the second (If the distance between the clustering center and the existing clustering center is less than the minimum distance between two cluster centers, the center is discarded, then the color with second most pixels is searched). Select in turn, until all Nc cluster centers are selected.

2.3 Extraction of Color Features

By clustering, a main color table of an image can be extracted. Representing the overall features of the whole image by using the main colors not only can effectively reduce the complexity of computing similarity, but also can effectively ensure the preservation of features after image processing. The characteristics we want to extract are mainly color and size.

Color feature: After clustering, the color characteristic of a cluster is represented by the average value of all pixels in the cluster. At the same time, in order to facilitate the next calculation of similarity, we transform the coordinates of h, s, v color space into Euclidean space, and use c1, c2, c3 to represent the transformation formula as follows:

$$c_1 = s\cos(h), \quad c_2 = s\sin(h), \quad c_3 = v \tag{1}$$

Size feature: We use ρ to represent size characteristics:

ρ= the number of pixels contained in clustering/Total number of pixels in an image.

For a plurality of main colors of an image, we use f to represent its features, a total of four dimensions: $f_i = (c_{i1}, c_{i2}, c_{i3}, \rho_i)$

The characteristic vector of the main color of an image can be expressed as $\{f_1, f_2, \ldots, f_n\}$, where *n* represents that the image has *n* main colors.

2.4 Calculate Image Similarity

Because images has different number of primary colors, we cannot simply take the distance between the main colors as the calculation element to detect their similarity. Assume there are two images to be detected, X and Y, whose main colors are m and n respectively. We calculate the similarity between the two main colors of the two images, so as to calculate the similarity of the whole image. The following formula is used to calculate the similarity between the two main colors:

$$\mathrm{a(i,j)} = 1 - 1/\sqrt{5}\left[(c_{i1} - c_{j1})^2 + (c_{i2} - c_{j2})^2 + (c_{i3} - c_{j3})^2\right]^{1/2} \tag{2}$$

In the above, *i* and *j* represent one of the main colors between the two images of *X* and *Y*, respectively. Through the similarity of main colors, we can easily calculate the similarity of image *X* to image *Y*.

$$\mathrm{S(X,Y)} = \sum_{i=1}^{m} W_i\, a\big(i, P_Y(i)\big)$$
$$\text{where } \sum_{i=1}^{m} W_i = 1 \tag{3}$$

In the formula, W_i represents the weight of the primary color *i*, which is determined by the number of pixels that the color occupies in the image, and represents the importance of the color in the image. For simplicity, its value can be determined by the assumption that the weight of the main color is proportional to the number of pixels. $P_Y(i)$ represents the mapping of *X* to *Y*, returning the i^{th} closest and most similar color in these two images.

3. EXPERIMENTAL RESULTS

We divide 256 colors of *r*, *g* and *b* into 16 parts by using the similar algorithm as statistical color histogram. Because the three color components are calculated together, the total number of colors is 16*16*16=4096. Next, we assign all pixels of an image to these color segments, and arrange them in descending order from large to small. The first 100 most frequently selected colors (segments) are selected as samples for clustering. Next, the 100 samples are clustered by ISODATA clustering algorithm. In the clustering algorithm, we select six initial clustering centers; the upper limit of standard deviation is 5; the minimum distance between two clustering centers is 30; the maximum number of merged clusters allowed in each iteration is 3; and the minimum number of samples allowed in each clustering is 10.

The samples are stored in the form of two-dimensional matrix, at the same time, each sample has three eigenvalues, representing *r*, *g*, and *b*. In calculating the distance from the sample to the cluster center and the distance between the cluster center, we calculate these three eigenvalues simultaneously. The average color in the cluster is calculated to represent the color of the image in the clustering. Before calculating the similarity, we convert the three colors of *r*, *g*, *b* into the colors of *h*, *s*, *v*. After conversion, the colors obtained are used for similarity calculation.

In calculating similarity, set the weight of different primary colors, which is generally determined by the proportion of the color in the

image. Therefore, the weight of different colors is generally different, and the value will change according to different images. In this experiment, in order to simplify the operation, we uniformly take the same weight for all the main colors, which may reduce the accuracy of the final detection of image similarity. It is worth mentioning that, the image similarity calculated by the algorithm will in general have two different results, the similarity of one image relative to the other, so we take the average of the two images as the final overall similarity of the two images.

The experimental results are shown in the figures. In Fig. 2, it can be seen that by rotating the same image, the detection result of similarity is 1, which shows that the algorithm has rotation invariance. For an image itself as shown in Fig. 3, the similarity is 1. That is, for the same image, the system can get the same primary color table, and then carry out similarity calculation and matching. Since the main color tables of the two images are consistent, so the result of similarity is 1.

Figure 2. Image and its rotated image (Similarity=100%)

Figure 3. Detection of the same image (Similarity=100%)

(Similarity=89.70%)

Figure 4. Similarity detection of different images

The similarity detection based on the main color does not consider the spatial information of the image. The advantage of this method is that it does not need to segment the image, and does not need to process the image according to the block, which greatly reduces the amount of calculation and speeds up the efficiency of the algorithm. However, like color histogram, this method will affect the accuracy of similarity detection to a certain extent. At the same time, because the main color is obtained by clustering, the quality of clustering algorithm directly affects the whole image eigenvalue, and then affects the final detection results. Therefore, a suitable clustering algorithm is particularly important in the system. At the same time, in the whole process of calculation, there are also many parameter setting, and these parameters can directly affect the final calculation results. Choosing the right parameters is very important. In order to find the right parameters to adapt to different types of images, a large number of experiments and judgments are needed.

4. CONCLUSIONS

Content-based image retrieval technology has always been one of the hot research topics, and image similarity detection is the most important aspect. Compared with traditional text-based retrieval, content-based retrieval uses various information contained in the image itself, such as color, texture, spatial information and so on, to achieve image matching. At the same time, this is also very difficult, because computer display and human visual senses are not always the same. In order to solve the problem of large amount of computation in similarity detection, we proposed a similarity detection method based on the main color table, which uses a few of the most colors in an image to represent the whole image, greatly reducing the amount of computation, and achieves reasonable results. However, this method has shortcomings needed to improved, such as losing the spatial information of image. So we can segment the image to get the appropriate image block, process the image according to the block, or we can decompose the main color matrix of clustering by singular value decomposition, and use the singular value as the eigenvalue of the image to detect.

Acknowledgements

This work was supported by the Shenzhen Science and Technology Innovation Committee with the grant number JCYJ20170306170559215 in China.

5. REFERENCES

[1] K. Otsuka, T. Horikoshi, and H. Kojima, Memory-based forecasting of complex natural patterns by retrieving similar image sequences, Proceedings 10th International Conference on Image Analysis and Processing, 1999, pp. 874-879.

[2] I. Ahamd, and Taek-Sueng Jang, Old fashion text-based image retrieval using FCA, Proceedings 2003 International Conference on Image Processing, 2003, Vol. 3, pp. 29-33.

[3] Juan Manuel Barrios, and Diego Diaz-Espinoza, Text-Based and Content-Based Image Retrieval on Flickr, Second International Workshop on Similarity Search and Applications, 2009, pp. 156-157.

[4] Burhan Ergen, and Muhammet Baykara, Content based medical image retrieval feature extraction of using statistical spatial methods for content based medical image retrieval, IEEE 18th Signal Processing and Communications Applications Conference, 2010, pp. 692-695.

[5] Hanife Kebapci, Berrin Yanikoglu, and Gozde Unal, Plant Image Retrieval Using Color, Shape and Texture Features,The Computer Journal, 2011, Vol. 54, Issue 9, pp. 1475-1490.

[6] Aun Irtaza, M. Arfan Jaffar, and Muhammad Tariq Mahmood, Semantic Image Retrieval in a Grid Computing Environment Using Support Vector Machines, The Computer Journal, 2014, Vol. 57, Issue 2, pp. 205-216.

[7] Deepak Ghimire, and Joonwhoan Lee, Color Image Enhancement in HSV Space Using Nonlinear Transfer Function and Neighborhood Dependent Approach with Preserving Details, Fourth Pacific-Rim Symposium on Image and Video Technology, 2010, pp. 422-426.

[8] Asmala Ahmad, and Suliadi Firdaus Sufahani, Analysis of Landsat 5 TM data of Malaysian land covers using ISODATA clustering technique, IEEE Asia-Pacific Conference on Applied Electromagnetics, 2012, pp. 92-97.

Quality Monitoring in Wire-Arc Additive Manufacturing Based on Spectrum

Yiting Guo, Zhuang zhao, Jing Han
Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense
Nanjing University of Science and Technology
Nanjing, China
1542313812@qq.com

Lianfa Bai
Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense
Nanjing University of Science and Technology
Nanjing, China
blf@njust.edu.cn

ABSTRACT
This paper presents a classification method based on spectrum for the quality monitoring in wire-arc additive manufacturing. Triggered by the field programmable gate array (FPGA), the spectrum was collected in the peak current. In this way, we acquired the spectrum with abundant information within one welding current period. In spectral analysis, we proposed two classification methods :priori threshold; KNN based on locality preserving projection (LPP-KNN). Spectrum can simultaneously evaluate the weld pool's status and monitor quality defects. Our method is not limited to one welding process, and experimental results of three wire materials in cold metal transfer (CMT) welding have verified the superiority of our method on the number of monitoring objects, accuracy and stability.

CCS Concepts
• Computing methodologies → Machine learning →Machine learning algorithms → Spectral methods.

Keywords
quality monitoring**;** spectral analysis; spectral classification; dimensionality reduction; metallurgical quality analysis.

1. INTRODUCTION
Addictive manufacturing is a developing technology which can reduce part cost and manufacture complex assemblies[1]. The combination of an electric arc as heat source and wire as feedstock is referred to as Wire-arc additive manufacturing (WAAM) [2]. On-line quality monitoring of the weld pool has been a research topic in welding field as well as a main content of the advanced manufacturing technology(AMT).This has become vital to the manufacturing industry for their survival and sustainability[3]. In traditional manual quality monitoring, skilled welders rely on visual information and personal experience mainly to monitor the welding quality [4,5]. In order to simulate the welder's judgement of weld pool quality by observing the weld pool, many researchers have adopted the vision sensor to capture weld pool images [6], which contain abundant information and are closely related to the welding quality. Currently, the visual quality monitoring around the world is widely based on passive vision.

As to the welding quality monitoring, Although weld pool images can be used to evaluate the weld pool's status, which is very intuitive in vision, this method cannot solve problems related to the changes of material composition in welding process--the accurate shielding gas flow rate and the slag inclusion, for example. In welding, the shielding gas flow rate can influence fluidity of the weld pool and welding quality; besides, some precision machining, such as nuclear industry, has a strict requirement for the processing environment. In this area, we must ensure that the environment is free from impurities. So it is necessary to monitor the shielding gas flow rate and the interference of impurities. Considering the week point of vision, other measurement data is needed to used for this question. This paper used the arc spectrum to solve this problem. The arc spectrum, which includes abundant information about the weld pool characteristics, is from the excitation of the shielding gas ionizing, the welding wire gasification, etc. On-line spectral analysis system can simultaneously measure multi channels and multi components, with faster speed and higher accuracy. In arc additive manufacturing, because of the unaltered material components, each layer of the spectrum shows little difference.

In order to obtain abundant information, we used the field programmable gate array (FPGA) to trigger the spectrometer in the peak current within one welding current period.

As to arc spectral analysis, Shea et al. [7] monitored the intensity ratio of the hydrogen line in 656.28 nm and argon line in 696.54 nm to measure the 0.25 % hydrogen in shielding gas. Mirapeix et al. [8] used a subpixel algorithm to calculate the plasma electronic temperature. This method can detect common defects in the welding seam caused by insufficient shielding gas flux or current fluctuations of the welding power source. Li et al. [9] detected some welding defects by using suitable spectral processing zones (250-300 nm, 750-830 nm, 776.6-777.6 nm, 867.5-868.5 nm, 900-1000 nm). García-Allende et al. [10] used sequential forward floating selection (SFFS) algorithm to realize spectral dimensionality reduction, and an artificial neutral network to carry out identification task. Zhang et al. [11] collected statistic characteristic parameters of certain spectrum bands of interest, adopted wavelet packet transform (WPT) to remove the pulse interference in the monitoring curves and used the signal-to-noise ratio (SNR) to monitor defects. In this paper, considering the quality monitoring contents and the spectral curve features, we adopted a method which combines the prior threshold with the k-

ICVIP 2018, December 29–31, 2018, Hong Kong, Hong Kong

ACM ISBN 978-1-4503-6613-7/18/12…$15.00
https://doi.org/10.1145/3301506.3301534

nearest neighbor (KNN) based on locality preserving projection (LPP-KNN) method to process spectral data.

We used two spectral methods, i.e. prior threshold-based method and LPP-KNN method. This method is not limited to one welding

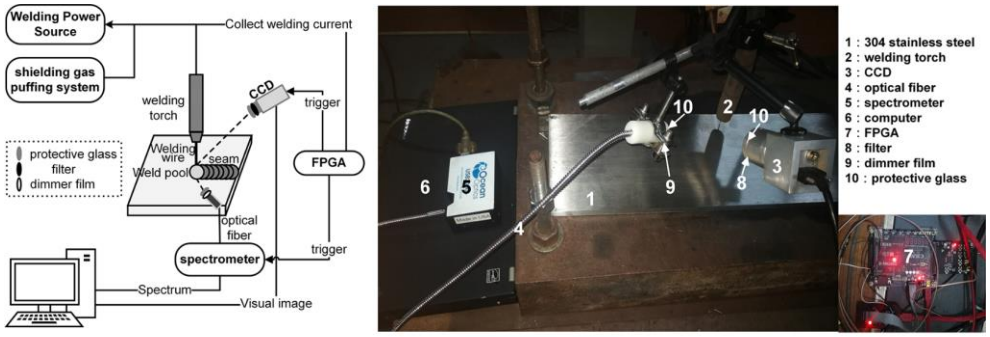


Figure. 1 Schematic illustration and diagram of the method based on spectrum

process. In this paper, we used the cold metal transfer (CMT) welding process as an example, and verified the superiority of our method with the experimental results in the welding of stainless steel, high-strength steel and high-nitrogen steel.

2. SYSTEM OF SPECTRUM

In this paper, we introduced a system of spectrum. Triggered by the FPGA, the spectral data was collected in the peak current. At the same time, the weld pool image was also captured in the base current, which is for observing more intuitionistic. In this paper, it is not described with a lot of words about camera.

The schematic illustration and diagram of the method of spectrum is shown as Figure. 1. The system consists of three parts: FPGA device, data collecting device, and data storage and display device. FPGA device includes FPGA development board (ALINX XILINX spartan-6 XC6SLX9); the data collecting device includes: protective glass plate, 850 nm high pass filter, 10%neutral dimmer film, CCD camera (BASLER acA1920-155 um), spectrometer (ocean optics USB2000+); and the data is stored and displayed on a computer.

During collecting spectrum, the working band of the spectrometer is 200-1100 nm, which suits the requirements of the quality monitoring.

This system is applicable to a variety of welding processes. In this paper, we used the CMT as an example, and the triggering time of the spectrum is shown as Figure. 2:

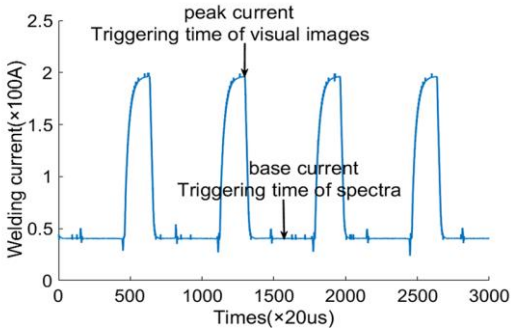
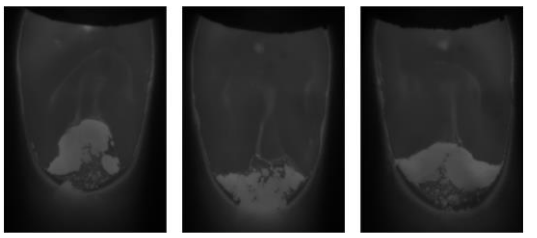


Figure. 2 Electrical signal diagram

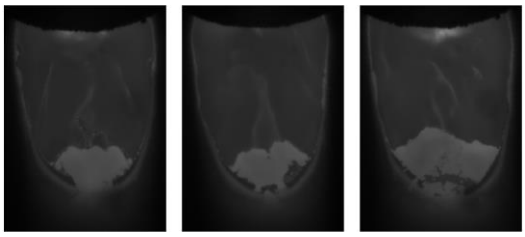
As to the spectral data, since the arc spectrum is the radiation caused by the current excite the shielding gas and metal vapor, it is better to collect the spectrum in the peak current at which the arc is the strongest.

3. QUALITY MONITORING

Take the stainless steel at the first layer of the single-pass multi-layer welding, 121A welding current and 30 cm/min welding speed as an example, the discrimination of the weld pool images at 25 L/min and 5 L/min shielding gas flow rates is not obvious as shown in Figure. 3(a), (b); the discrimination of the weld pool images with or without rust and oil stain on the base metal at 25 L/min shielding gas flow rate is not obvious as shown in Figure. 3(c), (d), (e), (f). The spectrum contains information about the material composition. Thereby the quality monitoring of shielding gas flow rate and residues on the base metal ,such as oil stain and rust both, rely on spectral analysis.



(a) normal gas **(b) low gas** **(c) no rust**



(d) rust **(e) no oil stain** **(f) oil stain**

Figure. 3 Weld pool images with different defects

We used two methods to analyze the spectral data:

1. As to the quality monitoring of spectrum that contains new material components, we analyzed the abnormal bands according to the atomic emission spectrum, selected specific bands and used threshold for the relative intensity of the spectrum or the ratio of the relative intensity to achieve quality monitoring;

2. As to the spectral curve, we first selected characteristic bands, then adopted the locality preserving projection (LPP) algorithm [12] to further analyze the spectral information and used KNN to classify the final spectral features.

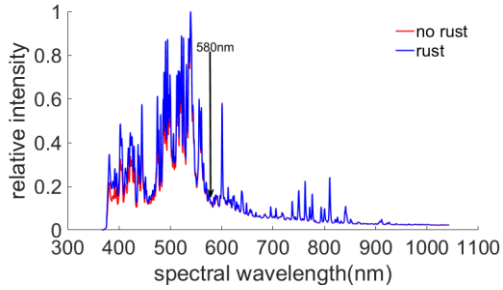
3.1 New Material Composition

New material composition is mainly referred to as doping here. Doping quality monitoring can greatly reduce welding quality problems, such as poor weld formation, porous, rough surface, etc. These welding problems are caused by residual oil and rust due to improper cleaning and grinding during welding process. It is widely used in nuclear industry and ASME shipbuilding industry.

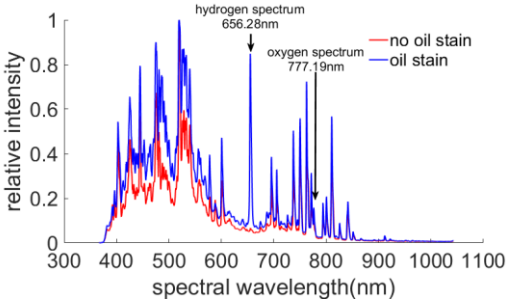
As to the doping quality monitoring of the spectrum, we analyzed the specific spectral curve and then selected representative bands and used the examples, such as base metal with or without rust and oil stain, to evaluate reliability of this method.

Iron oxide can form strong iron bands in 300-580 nm visible light [13]. We used the stainless steel at 25 L/min shielding gas flow rate, 30cm/min welding speed and 121A welding current as an example. The spectrum curve with or without rust is shown as Figure. 4(a). In the iron band, the relative intensity of spectrum with rust (Fe_2O_3) is higher than that without rust. Thereby, we applied a threshold for the average relative intensity of the normalized spectral data between 300-580 nm. If the average relative intensity is greater than the threshold for several times, it is concluded that rust exists.

The base metal with oil stain differs significantly from that without oil stain in the hydrogen spectrum. We used stainless steel at the first layer of the single-pass multi-layer welding, 25 L/min shielding gas flow rate, 30cm/min welding speed and 121A welding current as an example and the spectral curve is shown as Figure. 4 (b). The spectrum with oil stain displays obvious peaks in the hydrogen spectrum (656.28 nm). Besides, since the shielding gas contains oxygen, we concluded that the relative intensity of the oxygen spectrum (777.19 nm) remained the same. Thereby, we applied a threshold for the ratio of the relative intensity of the hydrogen spectrum and oxygen spectrum. If the average relative intensity is greater than the threshold for several times, it is concluded that oil stain exists.



(a) with or without rust



(b) with or without oil stain

Figure. 4 Spectrum with different defects

3.2 Shielding Gas Flow Rate

LPP-KNN classification, which is mainly used to monitor the shielding gas flow rate, which directly affect welding quality such as forming and surface roughness, uses the LPP dimensionality reduction to further analyze the characteristic bands and collect more spectral information, and then KNN is used to classify the final spectral features.

Take the stainless steel welding wire at the first layer of the single-pass multi-layer welding, 30 cm/min welding speed and 121A welding current as an example, and the spectrum at different shielding gas flow rates is shown as Figure. 5. At different shielding gas flow rates, the material components remain unaltered. Thereby the first kind of spectrum processing method, which is used in the quality monitoring of new material components such as oil stain and rust, is not suitable for the shielding gas flow rate monitoring. If applied, the robustness will descend. When processing the spectrum, the LPP-KNN algorithm displays strengthened adaptability, increased stability in monitoring and stronger robustness while at the same time prevents the inaccurate classification caused by unsuitable threshold and current interference.

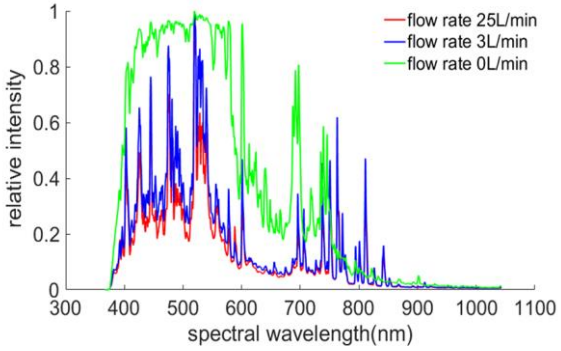


Figure. 5 Spectrum at different shielding gas flow rates

The LPP algorithm is based on the linear approximation of Laplacian Eigen maps [14]. The LPP algorithm, which is easy to use and fast in processing, can maintain the structure of local neighborhood of the samples in space when performing data dimensionality reduction.

Suppose the number of n dimension sample sets $\mathbf{X}=\{x_1,x_2,\mathrm{L}\ ,x_m\}$, where $x_i\in\mathbf{R}^n(i=1,\mathrm{L}\ ,m)$, is m, and the goal of the LPP algorithm is to use the projection matrix $\mathbf{W}=\{w_1,w_2,\mathrm{L}\ ,w_l\},l<\mathrm{n}$ to project the data sets $\mathbf{X}$ onto a low-dimensional feature space $\mathbf{Y}=\{y_1,y_2,\mathrm{L}\ ,y_m\}$, where $y_i\in R^l(i=1,\mathrm{L}\ ,m)$, i.e. $y_i=\mathbf{W}^Tx_i$. LPP algorithm identifies the optimal projection direction by minimizing the objective function.

$$\min\sum_{i,j}\left\|y_i-y_j\right\|^2\mathbf{S}_{ij} \tag{1}$$

We used the Kernel-weighted method to obtain the weight matrix $\mathbf{S}$, and the equation is:

$$\mathbf{S}_{i,j}=\begin{cases}\exp(-\left\|x_i-x_j\right\|^2/t) & x_i\ and\ x_j\ are\ neighbors\\ 0 & otherwise\end{cases} \tag{2}$$

The function can be simplified into generalized eigenvalue problem as:

$$\mathbf{XLX}^T w = \lambda \mathbf{XDX}^T w \tag{3}$$

where $\mathbf{D}$ is the diagonal matrix, and the elements on the diagonal line is the sum of the rows (or the columns) of the weight matrix, i.e. $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{S}_{ij}$. $\mathbf{L}$ is the Laplacian symmetric semidefinite matrix, $\mathbf{L} = \mathbf{D} - \mathbf{S}$. w is the eigenvector, λ is the eigenvalue.

After LPP dimensionality reduction, we used KNN classifier to perform shielding gas flow rate monitoring.

4. EXPERIMENTS AND RESULTS

4.1 Experimental Steup

Experimental devices include: an intelligent welding robot; FRONIUS CMT Advanced 4000Rnc welding power source; shielding gas puffing system; welding bench; stainless steel, high-strength steel, and high-nitrogen steel welding wires; 304 stainless steel base metal.

Experimental conditions of quality monitoring are shown as Table1:

Table 1 Experimental condition of quality monitoring

Welding method	process	current
single-pass multi-layer welding	CMT	Normal

For the stainless steel welding wire, the normal welding current is 121A; for the high-strength steel and the high-nitrogen steel, the normal welding current is 130A. The shielding gas consists of 98.5% argon and 1.5% oxygen. The total sample sizes in stainless steel, high-strength steel and high-nitrogen steel welding wires experiments are 9800, 13300 and 12200, respectively. 30% of these total samples are randomly sampled as train samples. These samples were collected at different times, which thereby are representative and can verify the stability of our method.

4.2 Results of Shielding Gas Flow Rate

Table 2 is the experimental results of three welding wires at different shielding gas flow rates and under different data dimensionality reduction methods. The shielding gas flow rates are 0 L/min, 3 L/min, 25 L/min. The three data dimensionality reduction methods are: Linear Discriminant Analysis (LDA) [15], LPP and Multi-Manifold Discriminant Analysis (MMDA) [16]. As shown in Table 3, LPP-KNN achieve best performance.

Table 2 Experimental results at different shielding gas flow rates

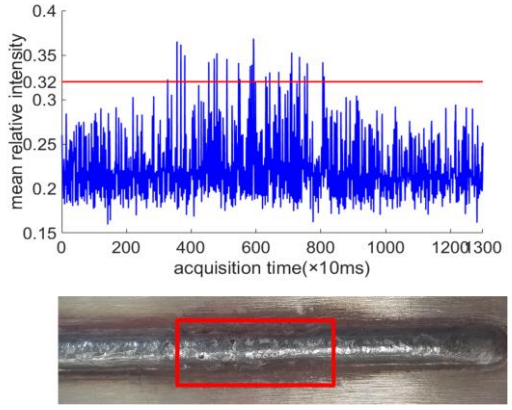
Welding wire	stainless steel(%)	high-strength steel (%)	high-nitrogen steel(%)
LDA-KNN	81.81	92.24	88.95
LPP-KNN	**98.90**	**96.85**	**99.95**
MMDA-KNN	81.83	32.68	81.59

4.3 Experimental Results of New Material Composition

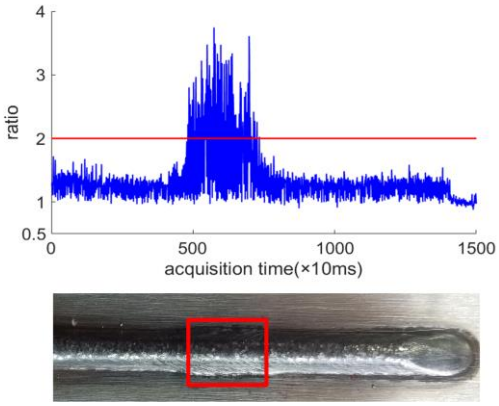
In order to verify the validity of the spectral method for monitoring new substance composition, we use oil stain and rust test.

As to the normalized spectral data, we applied a threshold for the average relative intensity in strong iron bands (300-580 nm) of visible light, and then detected the rust on the base metal. Figure. 6(a) is the analysis result of stainless steel base metal in single-pass multi-layer welding in which the middle part is spread with pure rust powder and the appearance of weld bead with rust defect. After several experiments, we set the threshold to 0.32. When the ratio is larger than 0.32 for three times, it is concluded that rust exists in the base metal.

We applied a threshold for the ratio of the relative intensity of hydrogen (656.28 nm) characteristic spectrum and oxygen (777.19 nm) characteristic spectrum and then detected the oil stain on the base metal. Figure 6(b) is the analysis result of the spectral data of base metal in single-pass multi-layer welding with oil stain in the middle part and the appearance of weld bead with oil stain defect. After several experiments, we set the threshold to 2. If the ratio is larger than 2 for five consecutive times, it is concluded that oil stain exists in the base metal.



(a) base metal with or without rust



(b) base metal with or without oil stain

Figure. 6 Experimental results of different defects and welding appearance

Compared with the methods of machine learning or neural network in some literatures, which is about quality monitoring of new material composition, this method is simpler and easier to explain.

5. CONCLUSION

In this paper, we proposed a method of spectrum. Triggered by the FPGA, the spectrum was collected in the peak current within one

welding current period. In this way, we obtained the spectrum with abundant information.

The spectral analysis of two classification methods, including priori threshold and LPP-KNN, can effectively perform quality monitoring of the weld pool. This method solves the short board of the monitoring the changes of material composition in vision. The experimental results of stainless steel, high-strength steel, and high-nitrogen steel based on CMT welding process have verified that this online quality monitoring of the weld pool can detect more quality problems and has higher recognition.

Limited by the actual band range and the sensitivity of the spectrometer, further research on the quality monitoring of spectrum with a small amount of doping is needed to improve the recognition rate. Besides, since the welding method in this paper is limited to the surfacing welding, other welding methods such as fillet welding need to be further researched.

6. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (no. 61727802 and 61501235).

7. REFERENCES

[1] Huang, J. G., Qin, Q., Wang, J., and Fang, H. 2018. Two Dimensional Laser Galvanometer Scanning Technology for Additive Manufacturing. *International Journal of Materials Mechanics and Manufacturing*, 6(5),332-336.

[2] Williams, S. W., Martina, F., Addison, A. C., Ding, J., Pardal, G., and Colegrove, P. 2015. Wire + arc additive manufacturing. *Materials Science and Technology* (7), 641-647.

[3] Sangster, N., Duke, R., Lalla, T., Persad, P., and Ameerali, A. 2016. Investigating the use of advanced manufacturing technologies in the manufacturing assembly sector in a small developing country. *International Journal of Materials Mechanics and Manufacturing*, 4(4),266-272.

[4] Liu, Y. K., Zhang, Y. M., and Kvidahl, L. 2014. Skilled human welder intelligence modeling and control: part 1 – modeling. *Welding Journal, 93*(5).

[5] Kovacevic, R., and Zhang, Y. M. 1995. Machine vision recognition of weld pool in gas tungsten arc welding. *Proceedings of the Institution of Mechanical Engineers Part B Journal of Engineering Manufacture, 209*(22), 141-152.

[6] Guo, B., Shi, Y., Yu, G., Liang, B., and Wang, K. 2016. Weld deviation detection based on wide dynamic range vision sensor in mag welding process. *International Journal of Advanced Manufacturing Technology, 87*(9-12), 3397-3410.

[7] Shea, J. E., and Gardner, C. S. 1983. Spectroscopic measurement of hydrogen contamination in weld arc plasmas. *Journal of Applied Physics, 54*(9), 4928-4938.

[8] Mirapeix, J., Cobo, A., and Conde, O. M. 2006. Real-time arc welding defect detection technique by means of plasma spectrum optical analysis. *Ndt & E International, 39*(5), 356-360.

[9] Li, Z., Wang, B., and Ding, J. 2009. Detection of gta welding quality and disturbance factors with spectral signal of arc light. *Journal of Materials Processing Technology, 209*(10), 4867-4873.

[10] García-Allende, P. B., Mirapeix, J., Conde, O. M., Cobo, A., and López-Higuera, J. M. 2009. Spectral processing technique based on feature selection and artificial neural networks for arc-welding quality monitoring. *Ndt & E International, 42*(1), 56-63.

[11] Zhang, Z., Yu, H., Lv, N., and Chen, S. 2013. Real-time defect detection in pulsed gtaw of al alloys through on-line spectroscopy. *Journal of Materials Processing Technology, 213*(7), 1146-1156.

[12] He, X. 2003. Locality preserving projections. *Advances in Neural Information Processing Systems, 16*(1), 186-197.

[13] He, T., Wang, J., Chen, Y., and Lin, Z. J. 2006. Study on spectral features of soil fe_2o_3. *Geography and Geo-Information Science, 22*(2), 30-34.

[14] He, X., Yan, S., Hu, Y., Niyogi, P., and Zhang, H. J. 2005. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol.27, pp.328--340).

[15] Zheng, W. S., Lai, J. H., and Li, S. Z. 2008. 1d-lda vs. 2d-lda: when is vector-based linear discriminant analysis better than matrix-based?. *Pattern Recognition, 41*(7), 2156-2172.

[16] Yang, W., Sun, C., and Zhang, L. 2011. A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognition, 44*(8), 1649-1657.