Emotion Estimation in Crowds: A Survey

Oscar J. Urizar[1,2], Emilia I. Barakova[2], Lucio Marcenaro[1], Carlo S. Regazzoni[1,3], Matthias Rauterberg[2]

[1]Department of Electrical, Electronic, Telecommunications
Engineering and Naval Architecture (DITEN)
University of Genoa - Genoa, Italy

[2]Department of Industrial Design
Eindhoven University of Technology
Eindhoven, Netherlands

[3]Carlos III University of Madrid

Keywords: Crowd emotions survey, collective emotions, emotion estimation, affective models, affective datasets.

Abstract

Emotions play an important role in human behavior, even more so in large congregations of people where emotional states are prompt to be contaged and amplified. This work presents a qualitative systematic review of the literature concerning the estimation of emotions and affects in real-life crowded environments, covering the aspects of methods and datasets. The academic search engine *Scopus* was inquired and the search was limited to publications in the English language addressing any of the aforementioned aspects. The aim of this contribution is to highlight advances, limitation and trends in addressing the estimation of emotions in crowds.

1 Introduction

Emotion is a concept centered on the individual, considered as a process directed to specific events, involving the appraisal of intrinsic features, with influence in multiple bodily systems and a strong impact on behavior; In this sense, collective (crowd) emotions occur when the same event yields a similar appraisal and elicits a common emotion among the members of a crowd. As emotions have a significant influence in the behavior of individuals, they are also essential in understanding the dynamics of a whole crowd. Affective computing (AC) delves in transferring the theoretical knowledge of emotions and affects into systems capable of recognize, model and express such human aspects. In this direction, significant advances have taken place in recent years, mainly focusing on the detection of emotions on single individual, from facial expressions to body language, speech analysis and physiological signals. Further research in the field of AC has started to address the emotions of individuals in groups and crowds; however some of the approaches intended for estimating emotions in individuals are not transferable to groups, emphasizing the need for inventive ways to address this issue. Several issues are introduced when shifting from the estimation of single to collective emotions: which behavioral cues and what sensors are appropriate for inferring emotions in crowds? does a microscopic or a macroscopic level of description better suited? is the emotion of a crowd the sum of the member's emotion or is it something else? to name a few. As cities grow larger and crowds become commonplace, systems capable to identify collective emotions are essential to optimize public spaces, manage crowd flows and ensure safeness.

The academic search engine *Scopus* was inquired, limiting the search to publications in the English language and using the combination of keywords: '*crowd* & *emotion*', '*collective* & *emotion*', '*social* & *emotion*', '*group* & *emotion*', '*crowd* & *affect*', '*collective* & *affect*', '*social* & *affect*', '*group* & *affect*', '*crowd* & *dataset*' and '*emotion* & *dataset*'. The focus of this qualitative systematic review is on literature dealing with physical crowds, defined here as real-life human groups of any size concurring at a physical location for a significant amount of time. Hence, publications addressing simulated or virtual crowds (e.g. online games/communities, social networks, blogs, etc.) are not considered here. Two main aspects of emotion estimation in crowds are addressed in this paper, namely methods and datasets. Methods are grouped into those dealing with emotion regulation and contagion, and those addressing the estimation of emotions. A representative selection of datasets intended for affective and crowd analysis tasks are examined and a discussion on their applicability to emotion estimation in crowded environments is provided. The aim of this contribution is to examine proposed methods to estimate emotions in crowds, highlighting their capabilities and limitations, and the currently available datasets for such purpose.

The remaining content is organized as follow: Methods applicable for emotion estimation in crowded settings are discussed in section 2. In section 3, datasets and benchmarks available for the evaluation of computational models are presented. Finally, section 4 provides conclusions on the examined methods and datasets along with a discussion on the advances, trends and shortcomings found in the inquired literature.

2 Methods

Several methods addressing emotions have been proposed, some simulate the emotional behavior of crowds, whereas others aim to analyze and understand it from real-world measurements. This survey focuses on the later kind of models, excluding methods that are not applicable to real crowds and focus purely on simulations.

2.1 Emotion Regulation and Contagion

An important aspect in understanding the dynamics of emotions in groups and crowds concerns the problem of whether and how these spread or amplify across individuals in a group. Inspired on neural mechanisms revealed by the recent work of Damasio [1], the authors of [2] and [3] propose ASCRIBE, an agent-based model to describe the interplay of mental states (*emotions*, *beliefs* and *intentions*) of individuals in the decision-making process under stressful situations. ASCRIBE is defined as having an external and internal level of operation; at the external level it incorporates mechanisms for mirroring mental states between individuals, at the internal level it describes how emotions and beliefs affect each other and how both affect a person's intentions. The model was put to the test by simulations and an empirical study case that compared four models and showed the ASCRIBE model to yield higher prediction accuracy. Expanding the concepts applied in the ASCRIBE model, the multi-agent-based model presented in [4] formalizes several concepts of emotion contagion spirals based on fundamental aspects at the individual level: the senders current emotional state and the extent to which the sender expresses the emotion; the strength of the communication channel from sender to receiver; the receiver current emotional state, its openness or sensitivity for the received emotion, bias to adapt emotions upward or downward and tendency to amplify emotions. Although no empiric validation is provided, the model is tested with simulations and mathematical analysis, and it produced interesting emerging patterns identified in psychology literature such as the upward and downward emotion spirals. Further studying the role of emotions in the decision-making process under stressful situations, and with similar concepts to ASCRIBE, an adaptive agent model for affective social decision making is proposed by Manzoor et al in [5] and later extended in [6] to account for emotion regulation and contagion. This model incorporates Hebbian learning principles to adapt the agent's decision-making process, but as found in the experiments conducted, it did not yield significant discrimination in the agent's decisions. Regulation is approached by antecedent-focused strategy (regulation before an emotional response has an effect on behavior), modeling it as a dynamic interaction between internal mental states and contagion is implemented as described in [4].

2.2 Emotion Estimation

Focusing on the task of emotion estimation, the framework introduced in [7] addressed the recognition of individuals' membership and emotions within a group setting by means of multimodal analysis of facial and body expressions. Faces are represented by facial landmark trajectories and extended volume Quantised Local Zernike Moments (QLZM) [8], and encoded into Fisher Vector (FV) [9] representations as input to a Gaussian Mixture Model (GMM) classifier to recognize emotions in arousal and valence dimensions. The framework was tested with a self-collected dataset of 3 groups of 4 individuals each, monitored while watching a movie. The proposed approach vQLZM-FV outperformed the compared methods, namely Facial landmarks, body HOG and body HOF. Although this approach focused on investigating the affective response of individuals while watching long-term videos, it is theoretically applicable to crowds under the assumptions that crowd members' face and body are visible for a long-enough interval and within an acceptable resolution.

The authors of [10] summarize their previous work on three bio-inspired probabilistic algorithms for perception of emotions from crowd dynamics. The first algorithm starts by partitioning the environment using an Instantaneous Topological Map (ITM) and a Dynamic Bayesian Network (DBN) is employed to model conditional interactions occurred in each subregion. these interactions are then converted into super states using a Self-organizing map (SOM) and the occurrence of these super states (events) are encoded by a Gausian mixture model as positive or negative emotions. The second algorithm starts with the detection of events, collecting them over time to obtain behavioral patterns which are then clustered into classes by means of a DBN; the distribution of these classes are modeled using GMM, building one model for positive and one for negative emotions, to finally detect the emotional state by a likelihood ratio test. In the third algorithm the trajectory of single individuals are expressed as transitions (events) between subregions using a DBN and separated models are constructed for the event sequences labeled with a positive or negative emotion, to conclude with a log-likelihood test to determine the emotion according to the movement pattern exhibit by the individual. All three approaches are tested under a simulated scenario, showing the third algorithm to yield the highest emotion prediction accuracy according to the experiments conducted.

The authors in [11] propose a hierarchical Bayesian model aiming to describe the crowd both at the microscopic and macroscopic level. This approach uses pedestrians trajectories to create a topological map by means of a self-organizing map (SOM), dividing the environment into zones. At the microscopic level, the pedestrians trajectories are described as a Markov process transitioning between zones, and behaviors are modeled according to the origin and destination; each behavior is assigned an emotional label (positive, neutral or negative) according to the time required to reach the estimated destination. At the macroscopic level, the crowd is described by a vector state counting the number of people in each zone at a given time, a second SOM clusters and reduce the dimension of the state vectors to describe the dynamics of the crowd as a Markov process; finally the emotion of the crowd is assigned to be the predominant one as displayed by single individuals. This method is validated by a simulated crowd under different levels of crowd density and multiple behaviors.

3 Datasets

A common test bed is essential to measure and compare performance among different methods. This section examines two types of datasets, those designed for affective states recognition and those intended for crowd analysis. The objective is to determine whether the reviewed datasets are suited to test computational models dealing with emotion estimation in crowds, for which a discussion is provided in section 4. The datasets considered in this section are not exhaustive but rather representative of the diversity available for such tasks. Both groups of affective and crowd analysis datasets are listed in tables 1 and 2 respectively.

3.1 Affective Datasets

This subsection considers datasets intended for all kinds of affective tasks, not limiting to those fitted for crowds and using different sensors and ground truth formats in order to provide a comprehensive view of the available options.

Delving in the task of detecting emotional states, facial expressions have become a popular choice due to their universality and intrinsic relation to emotions [12]. By means of conventional cameras and in a controlled environment, the datasets CMU [13] [14] and FER-2013 [15] collected static images of facial expressions from participants who were requested to act different emotional states following the discrete emotions scheme [16]. Aiming to simplify the collection of static images and to reach a greater number of participants, the authors of the Gamo dataset [17] made use of a web-based interface were participants play a game by performing specific facial expressions captured by a web camera. Progressing from static images only, the CK+ dataset [18] provides sequences of images where participants enact a series of facial expressions, and emotions are described in terms of facial action units [19].

Expanding the scope of behavioral markers, the dataset CREMA-D [20] contains short videos of participants displaying facial and vocal expressions for the study of multi-modal emotion expression and perception, whereas the dataset LIRIS-ACCEDE [21] [22] goes one step further and captures body expressions.

However, as the authors in [23] argue, using conventional 2D cameras lacks robustness as this kind of cameras are subject to poor illumination and changes in lighting conditions. In response, they propose the use of Kinetic cameras as these are able to capture depth, and produce a dataset containing 3D models of several participants performing multiple facial expressions. Moving from emotions (brief affective states) to moods (long term affective states), the work in [24] introduces the EMMA database which employs both 2D and kinetic cameras, and provides longer intervals of data capture as the dataset is intended for mood recognition.

Focusing on physiological measurements, the DEAP dataset presented in [25] [26] collected the electroencephalogram (EEG), electrooculogram (EOG), Galvanic skin response (GSR), blood volume pressure (BVP), temperature and respiration signals of participants. And a frontal video face was recorded for some of those participants. One-minute long excerpts of music videos were used as the stimulus to elicit emotions along the four quadrants of the arousal-valence plane.

3.2 Crowd Analysis Datasets

The fields of computer vision and crowd analysis are favored with an overgrowing importance and share a common interest in studying crowded environments, and as a result, multiple datasets have been produced. In compiling such datasets, cameras remain to be the preferred sensor for studying crowds due to the already widespread use of surveillance cameras in most public spaces.

Depending on the focus of study, datasets are designed to capture the desired circumstances. The popular dataset PETS 2009 [27] collected image sequences from multiple cameras with the aim to serve as a test bed for algorithms intended for people counting, density estimation, people tracking, flow analysis and event recognition. All the presented situations are mainly poor in terms of emotional behavior, except for the scenario S3 (event recognition) where an evacuation (rapid dispersion) is observed and can be associated to an emotional state of fear. The authors of CAD [28] recreated several normal collective behaviors adding the challenges of change in illumination and wavering trees in the background, however, the captured situations are not representative of any emotional behavior. Taking advantage of the large number of people attending the World Exposition of 2010 in Shanghai, the massive dataset Shanghai Expo 10 [29] was gathered. It provides a large amount of annotations at a regional level denoting crowd density, collectiveness and cohesiveness features under normal situations, but it lacks any relevance for inferring affective states. Focusing on groups and crowds, the authors of [30] present the Atomic Group Action dataset targeting the dynamics of group formation, yet no meaningful emotional behavior is exhibit. Rabiee's dataset [31] provides some emotional-rich situations such as panic and fight, although in a staged way. Finally, the S-hock dataset [32] focuses on the behavior of spectator crowds with rich annotations at the individual level, enabling the addition of further affective annotations although restricted to this type of crowds.

4 Discussion

In developing a well-grounded computational model, some theoretical issues regarding emotions in crowds must be addressed. One is to establish a working definition of what exactly is meant when talking about a crowd (e.g. features, properties) as there is no consensual definition in the fields of psychology and sociology. Another related aspect is to state if a method is limited to function for certain types of crowds, as these emerge in very diverse contexts and exhibiting a wide range of behaviors. Finally, it is important to provide a definition of what is meant by emotion and the emotional theory employed (e.g. discrete emotions, valence-arousal). Is the purpose of the method to estimate the emotion of individuals within a crowd or the emotion of the crowd as a whole? Except for [11],

Dataset	Modality	Sensory Data	Annotations	Naturalness
3D Face Model [23]	Facial expressions	Kinetic camera	Normal, happiness, sadness, surprise, anger	Acted
CK+ [18]	Facial Behavior	Image sequences	Anger, disgust, fear, happiness, sadness, surprise, contempt	Acted
CMU [13] [14]	Facial expressions	Static images	Happiness, sadness, anger, neutral	Acted
CREMA-D [20]	Facial and vocal expressions	Camera	Happiness, sadness, anger, fear, disgust, neutral	Acted
DEAP [25]	Facial expressions, physiological measurements	EEG, EOG, GSR, BVP, temperature, respiration	Valence, arousal, dominance, liking, familiarity	Induced
EMMA [24]	Facial and body expressions	Camera, kinetic camera	Valence, arousal	Induced and acted
FER-2013 [15]	Facial expressions	Static images	Happiness, sadness, anger, surprise, disgust, fear, neutral	Acted
GaMo [17]	Facial expressions	Static Images	Anger, disgust, fear, happiness, neutral, sad, surprise	Acted
LIRIS-ACCEDE [21]	Facial, vocal and body expressions	Camera	Valence, arousal	Acted

Table 1. Affective Datasets

Dataset	Modality	Sensory Data	Annotations	Naturalness
Shanghai Expo 10 [29]	Crowd movement	Camera	crowd density, collectiveness and cohesiveness	natural
Rabiee's [31]	Crowd movement	Camera	Panic, fight, congestion, obstacle, neutral behaviors	Acted
PETS 2009 [27]	Crowd movement	Image Sequences	Pedestrians' bounding box and location	Acted
CAD [28]	Crowd movement	Camera	Bottleneck, departure, lane, arch/ring and blocking crowd behaviors	Acted
S-Hock [32]	Individual behavior	Camera	People detection, head detection, head pose, body position, posture, locomotion, action/interaction, supported team, best action, social relation.	Natural
Atomic Group Actions [30]	Group actions	Camera	Group-group actions (formation, dispersal, movement) and group-person actions (person joining, person leaving)	Natural

Table 2. Crowd Analysis Datasets

the examined methods fail to provide a working definition of a crowd and implicitly indicate the type of crowd addressed by the method. Similarly, none provide a definition of emotion but the majority do indicate the emotional theory either implied or clearly stated. If a method is intended to estimate the emotion of individual crowd members then a definition as provided in the introduction of this paper is adequate; however, if a method aims to estimate the emotion of a crowd as a whole, a more clear definition is necessary.

The examined methods tend to model emotions in crowds either at the individual level (microscopic) or at a global (macroscopic) level. Microscopic models describe collectives at the individual level, modeling the emotions, behavior, actions and decisions of single crowd members. Macroscopic

models examine the crowd as a whole and describe it by means of global features, as a single entity evolving over time. On its own, both the microscopic and macroscopic models ignore important aspects of a crowd. Microscopic models are unable to capture the collective aspects of emotions in crowds, whereas macroscopic models fail to describe the interaction among individuals that foment the emergence of emotions. When capturing the affects of a crowd, an appropriate model should be able to depict: (a) the emotion of the individual, (b) the interaction between individuals, and (c) the crowd as a whole [33]. The majority of the examined models take a microscopic approach and focus on individual emotions of the members within the group or crowd, failing to capture the interaction of crowd members and the global essence of collective emotions. A common trait in the examined methods is the use of crowd members ambulatory behavior to infer emotional states. The reason to favor the use of ambulatory behavior over facial or body expressions when estimating emotions in crowds is that it enables the method's applicability to more (but not all) types of crowds, with different density levels and possibly limited visibility of the members face and body.

The choice of sensors and features to be used in order to estimate emotional states is central in discussing datasets suitable for crowds. Due to the challenging circumstances of crowded environments, noninvasive sensors such as surveillance cameras are preferred. Either at the microscopic or macroscopic level, the features used for individual emotion estimation are generally not suited for crowded environments due to multiple reasons: the faces and body of pedestrians are not always visible or can suffer from occlusion, and vocal expressions are easily distorted or impractical to perceive. Under these circumstances, visual information about the movement behavior of individuals becomes the most practical cue to infer emotional states but with the cost of higher uncertainty as the relation between behavior and emotion is highly dependent on the context of the situation. Given the above observations, all the examined affective datasets are rendered inadequate for methods devoted to detecting crowd emotions. The examined datasets intended for crowd analysis provide no annotations or meaningful behaviors in terms of emotions, only the S-hock dataset presents sufficient relevant affective information but is limited to spectator crowds. A dataset well suited for emotion estimation in crowds needs to capture diverse and meaningful behaviors accompanied with well validated affective annotations, ideally for multiple types of crowds and different emotions. The absence of publicly available datasets for emotion estimation in crowds is clearly evidenced in how the existing body of literature is evaluated. Throughout this survey, two main trends were identified: simulations and case-specific footage. Simulations are a practical solution for obtaining experimental data, however its arguable how well such simulations can replicate the complexity of emotional behavior. Case-specific footage is advantageous due to its naturalness but a more thorough evaluation in multiple cases is necessary to prove how well a method can be generalized. The absence of a common dataset prevents proposed methods to be properly evaluated and compared, decelerating further research in this area.

Acknowledgements

This work was partially supported by the Erasmus Mundus joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA, Agency of the European Commission under EMJD ICE. Prof. Carlo Regazzoni contributed to this paper during his stay at Carlos III University of Madrid as a visiting professor.

References

[1] A. R. Damasio. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 1996.

[2] M. Hoogendoorn, J. Treur, C. N. Van Der Wal, et al. Modelling the interplay of emotions, beliefs and intentions within collective decision making based on insights from social neuroscience. *Lecture Notes in Computer Science*, 2010.

[3] T. Bosse, M. Hoogendoorn, M. Klein, et al. Agent-based modelling of social emotional decision making in emergency situations. *Understanding Complex Systems*, 2013.

[4] T. Bosse, R. Duell, Z. A. Memon, et al. Agent-Based Modeling of Emotion Contagion in Groups. *Cognitive Computation*, 2014.

[5] A. Sharpanskykh and J. Treur. An adaptive agent model for affective social decision making. *Biologically Inspired Cognitive Architectures*, 2013.

[6] A. Manzoor and J. Treur. An agent-based model for integrated emotion regulation and contagion in socially affected decision making. *Biologically Inspired Cognitive Architectures*, 2015.

[7] W. Mou, H. Gunes, and I. Patras. Automatic Recognition of Emotions and Membership in Group Videos, 2016.

[8] E. Sariyanidi, V. Dali, S. C. Tek, et al. Local Zernike Moments: A new representation for face recognition. In *Proceedings - International Conference on Image Processing, ICIP*, pages 585–588, 2012.

[9] J. Sánchez, F. Perronnin, T. Mensink, et al. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 2013.

[10] M. W. Baig, M. S. Baig, V. Bastani, et al. Perception of emotions from crowd dynamics. *International Conference on Digital Signal Processing, DSP*, 2015.

[11] O. J. Urizar, M. S. Baig, E. I. Barakova, et al. A Hierarchical Bayesian Model for Crowd Emotions. *Frontiers in computational neuroscience*, 2016.

[12] P. Ekman. Facial expression and emotion, 1993.

[13] T. M. U. Mitchell. UCI Machine Learning Repository: CMU Face Images Data Set, 1999.

[14] D. Das and A. Chakrabarty. Emotion Recognition from Face Dataset Using Deep Neural Nets. *IET Computer Vision*, 2016.

[15] I. J. Goodfellow, D. Erhan, P. Luc Carrier, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 2015.

[16] P. Ekman. An argument for basic emotions, 1992.

[17] C. Tsangouri. Towards an In-the-Wild Emotion Dataset Using a Game-based Framework, 2016.

[18] P. Lucey, J. F. Cohn, T. Kanade, et al. The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotionspecied expression. *Cvprw*, 2010.

[19] P. Ekman and W. V. Friesen. The Facial Action Coding System. *Consulting*, 1982.

[20] H. Cao, D. Cooper, M. Keutmann, et al. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 2014.

[21] Y. b. Baveye, J.-N. Bettinelli, E. Dellandréa, et al. A large video data base for computational models of induced emotion. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pages 13–18, 2013.

[22] Y. Baveye, C. Chamaret, E. Dellandréa, et al. A Protocol for Cross-Validating Large Crowdsourced Data: The Case of the LIRIS-ACCEDE Affective Video Dataset. *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, 2014.

[23] S. Chickerur and K. Joshi. 3D face model dataset: Automatic detection of facial expressions and emotions for educational environments. *British Journal of Educational Technology*, 2015.

[24] C. Katsimerou. Crowdsourcing Empathetic Intelligence : The Case of the Annotation of EMMA Database for Emotion and Mood Recognition. *Acm Tist*, 2016.

[25] M. Soleymani, S. Member, and J.-s. Lee. DEAP : A Database for Emotion Analysis Using Physiological Signals, 2012.

[26] G. Placidi, P. Di Giamberardino, A. Petracca, et al. Classification of Emotional Signals from the DEAP dataset. *Proceedings of the 4th International Congress on Neurotechnology, Electronics and Informatics*, 2016.

[27] J. Ferryman and A. L. Ellis. Performance evaluation of crowd image analysis using the PETS2009 dataset. *Pattern Recognition Letters*, 2014.

[28] M. A. Hassan, A. S. Malik, W. Nicolas, et al. Reliability of bench-mark datasets for crowd analytic surveillance. *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*, 2015.

[29] C. Zhang, K. Kang, H. Li, et al. Data-Driven Crowd Understanding: A Baseline for a Large-Scale Crowd Dataset. *IEEE Transactions on Multimedia*, 2016.

[30] R. J. Sethi. Towards defining groups and crowds in video using the atomic group actions dataset. *Proceedings - International Conference on Image Processing, ICIP*, 2015.

[31] H. Rabiee. Novel Dataset for Fine-grained Abnormal Behavior Understanding in Crowd, 2016.

[32] F. Setti, D. Conigliaro, P. Rota, et al. The S-Hock dataset: A new benchmark for spectator crowd analysis. *Computer Vision and Image Understanding*, 2017.

[33] Cabinet Office. *Understanding Crowd Behaviours*. 2009.

[34] E. I. Barakova, R. Gorbunov, and M. Rauterberg. Automatic Interpretation of Affective Facial Expressions in the Context of Interpersonal Interaction. *IEEE Transactions on Human-Machine Systems*, 2015.

Infrared Thermography Processing to Characterize Emotional Stress: A Pilot Study

E. Gómez de Mariscal[1], A Muñoz-Barrutia[1], J. de Frutos[2], A. P. González-Marcos[3], A. M. Ugena Martínez[4]

[1] Bioengineering and Aerospace Engineering Department, Universidad Carlos III de Madrid,
Instituto de Investigación, Sanitaria Gregorio Marañón, Madrid, Spain
[2] Department of Physical Electronics, Electrical Engineering and Applied Physics, [3] Department of Photonic and Bioengineering Technology and [4] Department of Applied Mathematics,
ETSI de Telecomunicación, at the Universidad Politécnica of Madrid, Spain.

Keywords: Image processing, thermal imaging, classification methods, pattern recognition, emotional stress.

Abstract

Infrared Thermography (IRT) is a widely used technique for the detection of temperature change patterns. In the present study, we are interested in its application to the non-invasive follow-up of temperature changes on emotionally stressed people. Nine subjects performed the arithmetic task of the Trier Social Stress Test (TSST) while IRT videos were acquired. A pipeline was implemented to follow the temperature changes on Regions-of-Interest (ROIs) of the facial IRT videos. The pipeline is divided in four main blocks: (1) Frame selection, (2) image preprocessing (i.e., noise reduction), (3) face segmentation, (4) detection of ROIs (i.e., forehead, nose) and temperature extraction. Block 2 relies in an implementation of Growing Hierarchical Self-Organizing Maps (GHSOM). Faces are segmented (Block 3) with Hybrid Geodesic Region-Based Active Contours (HGRBAC). The detection of the ROIs (Block 4) is performed using the Viola-Jones classification method and finally, the temperature is extracted on the ROIs as a function of the time to enable the identification of stress patterns. Each subject's performance on the TSST tasks was measured by an Efficiency Ratio, being its value related to the level of concentration of the subject while performing the exercise.

1 Introduction

Nowadays, the tendency is that a technique to be incorporated into clinical diagnosis should be as non-invasive as possible. In the context of emotional states evaluation, in particular stress, vital signs are employed as markers (e.g., heart rate frequency and its variability, finger temperature, alpha-amylase levels, cortisol levels), [1]. Recently, the estimation of the face temperature acquired using Thermal Infrared cameras has been incorporated as a non-invasive subrogated stress measure [2]. The main goal of the present work is to provide a tool that facilitates the extraction of the stress markers from Infrared Thermography (IRT) images.

In the last decade, IRT has found a number of applications. In Guzman *et al.* [3], a robust image processing pipeline was presented for the extraction of facial blood vessels patterns to build the subject facial signature with application in biometrics. In the medical field, its utility to detect and monitor a variety of diseases has been demonstrated, between them, breast cancer [4], vascular diseases (in particular Raynaud's disease [5]) and arthritis [6].

Recently, several studies used IRT technology in an attempt to establish the relationship between emotions and the regulation of facial skin temperature. The fact that human body is thermally regulated and that the affective status of the individual alters some vital signs such as blood flow, lead us to the hypothesis that inner physiological reactions caused by emotions can be measured by thermography. The early work of Garbey *et al.* [7] presents a methodology to infer blood vessels location and estimate the blood flow from IRT static images. Later, Puri *et al.* [8] employed Garbey's approach to characterize the subject stress level using the temperature in the forehead skin as a subrogated marker. Jenkins *et al.* [9] performed a complete study to investigate IRT's potential to measure affective statuses (i.e., anxiety, relaxation, enthusiasm) changes. In this case, they compared the information extracted by a visual inspection of the IRT images, with the one given by an electroencephalogram (EEG) and a psychological test (Affective Self Report), but they did not develop any computational tool for the analysis. They showed that there exists a possitve correlation between the forehead temperature and the EEG activity, which implies that forehead temperature dynamics can be used to measure human internal status.

In this work, we adopt Puri *et al.* [8] approach and use the temperature of selected regions on the face as a subrogated marker for the stress level. The main contribution of the work is the development of a processing pipeline for the analysis of IRT videos. In particular, the pipeline is devoted to the accurate extraction of the skin temperature over time on desired facial Regions-of-Interest (ROIs), namely, whole face, forehead and nose.

The principal steps of the work which will be detailed in the subsequent sections, are presented in Figure 1. As we focus on the evaluation of emotional stress, the IRT videos were acquired while the subjects performed one of the arithmetic tasks

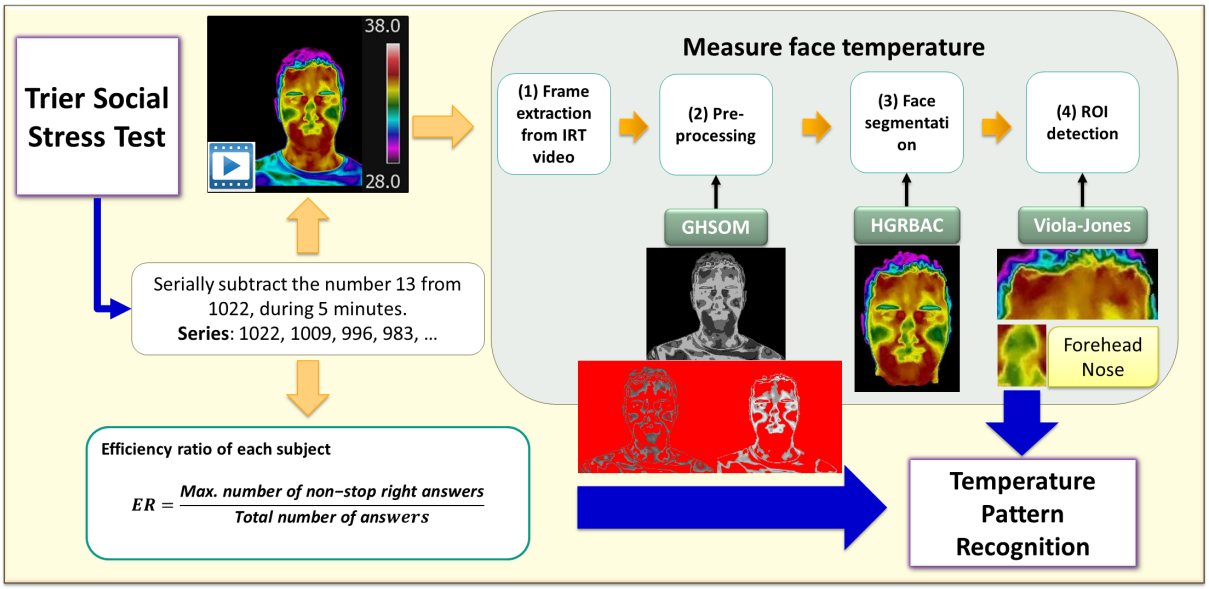


Figure 1. Illustration summarizing the main steps of the workflow. The pipeline is divided in mainly two blocks: (1) Computational method to measure the temperature and (2) Efficiency Ratio calculation. The computational methods consist on four blocks that combine techniques such as Growing Hierarchical Self-Organizing Maps (GHSOM), Hybrid Geodesic Region Based Active Contours (HGRBAC) and Viola-Jones detection algorithm. The ROIs to measure the temperature are face, forehead and nose.

defined by the Trier Social Stress Test (TSST) [10]. The results of the TSST serve to compute a measure of the mental concentration of the subject, called the Efficiency Ratio. This measure of concentration is compared with the temperature extracted on the facial ROIs and some conclusions of the subject stress level and their evolution while performing the task are extracted.

The structure of the manuscript is as follows: Section 2 describes the TSST task performed and the IRT video acquisition. Section 3 presents the pipeline implemented to extract the skin temperature over the ROIs. Section 4 describes the experimental results which are discussed in Section 5.

2 Materials

All the subjects of the study had to complete the arithmetical task of the TSST which consists of a serial subtraction of the number 13 from 1022, during 5 minutes (i.e., the resulting series is 1022, 1009, 996, 983, and so on). The stress test was applied to nine healthy volunteer students, four women and five men, within ages ranging from 22 to 32. The data employed for the analysis consists of IRT videos and images, both of them taken with an FLIR T640 infrared camera calibrated for a temperature in the range $[28°, 38°]$, see Figure 1. IRT images were taken at baseline and rest phases, and an IRT video of five minutes length was recorded during the arithmetical test. So, we had nine videos and 18 images to analyze.

Additionally, the whole experiment was recorded with a regular camera so as to collect the answers to the exercise for all the subjects.

The quantitative analysis of the IRT images and the videos was implemented in Matlab.

3 Methods

In this section, we describe the computational pipeline designed for the extraction of the ROIs from the IRT images. The workflow as illustrated in Figure 1, consists of four main steps: (1) Frame extraction, (2) preprocessing, (3) face segmentation and (4) ROI detection and temperature estimation.

3.1 Frame extraction

Instead of processing the whole IRT video, a number of frames were selected to lower the computation demand. Ten images were extracted per minute - one image every six seconds -, which results in 50 images. The initial (minute 0) and the final (minute 5) images were also kept. If we take into account the static images that we took at baseline and rest events, a total number of 54 images is selected per subject.

3.2 Preprocessing

IRT images are characterized by the presence of rough texture areas. In other words, there is a large variability in the local pixel intensities. Hence, noise reduction is needed previous to the face segmentation. Our proposal is to apply the Growing Hierarchical Self-Organizing Maps (GHSOM) [11] to quantize the image. The GHSOM provides an automatic discretization of pixel values in the image. When the pixel values are replaced by the classification obtained with the GHSOM, most of the noise in the image is removed and the convergence of the following level set segmentation is improved.

The unsupervised Self-Organizing Maps (SOMs), presented by Kohonen *et al.* in [12], are a type of Artificial Neural Networks of just one layer used for classification. SOMs constitute a dimensionality reduction method that produces a

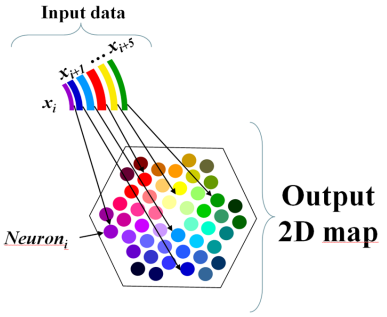


Figure 2. In every iteration i, weights between the neurons in the output layer and the input vector $\vec{x}_i$ are calculated. Each $\vec{x}_i$ will be associated to the neuron with the optimal weight. The neurons in the output layer will be located by the neighbourhood criteria, so two neurons will be close to each other if the data associated to each of them is similar (i.e., blue neurons are close to each other, red neurons are close to each other and between them, purple and pink are located).

two-dimensional discretized representation of the input space given by the training samples (a map, see Figure 2). SOMs apply competitive learning instead of error-correction learning as other neural networks. In particular, they use a neighborhood function to preserve the topological properties of the input space.

The Growing Hierarchical Self-Organizing Map (GHSOM) presented by Dittenbach *et al.* in[11], is an artificial neural network model with hierarchical architecture composed of independent growing self-organizing maps (layers). The improvements over the SOM architecture are twofold: The network is not static (i.e., the number of neurons in each SOM changes) and the capabilities to represent hierarchical relations between the data are more powerful. Figure 3 provides an intuitive explanation of how a GHSOM works.

Here, we propose to restore IRT images using a GHSOM classifier. The input data of each layer in the GHSOM is classified in the neurons of the layer based on its standard deviation. The number of neurons in the layer grows until the mean quantization error (i.e., the average of the mean error in each neuron) is less than a fix portion τ of the standard deviation. Likewise, all the neurons representing a portion of the mean quantization error higher than δ, will spread in a new layer (i.e., new SOM). We have tuned τ and δ to achieve a good compromise between the quantization level (i.e., classification depth) and the noise reduction. We trained the GHSOM classifier with the baseline image for each subject and applied it to the rest of his/her IRT images.

3.3 Face segmentation

IRT data provides complete frontal views of the subjects as shown in the workflow in Figure 1, so we need to compute a mask of the face. Guzman *et al.* [3] suggested the use of active contours to segment the face, so, in particular, we chose the Hybrid Geodesic Region-Based Active Contours (HGRBAC) [13] for this purpose.

The classical approach of active contours is based on deforming an initial contour towards the boundary of the object to be detected. The deformation is accomplished by minimizing a functional designed so that its local minimum is obtained at the desired boundary.

The HGRBAC is based on a gradient descent flow that is a combination of local geodesic active contours and more global region-based active contours [13].

Let $C(s)$ be the parametric curve evolving the object in the image I and s the Euclidean arc length parameter. Then, the geodesic active contours look for the local minimum length curve minimizing an energy functional E that depends on the length of the curve and its smoothness [14]. The Chan-Vese model [15] is a type of region-based active contour model in which the boundary is chosen so as to minimize the differences in the intensity values of the foreground and background regions. So, the energy functional of the HGRBAC is defined as the integration of the previous ones and is given by

$$E = \oint_{C(s)} \left(\int_{\Omega} (I_\chi - u_l)^2 dA + \int_{\bar{\Omega}} (I_\chi - v_l)^2 dA \right) ds. \quad . \quad (1)$$

where Ω and $\bar{\Omega}$ are the respective foreground and background defined in a local neighborhood of $C(s)$, and u_l and v_l are the mean intensity values of Ω and $\bar{\Omega}$, respectively. The characteristic function χ has value one inside the neighborhood and zero elsewhere.

The HGRBAC calculates the energy E iteratively and therefore it is necessary to tune the maximum number of iterations, the step for the integral and the size of the local neighborhoods. In this case, HGRBAC were initialized on the subject baseline image by a manual delineation of the face performed by the user. The output boundary of the active contours for a given time point served to initialize them for the IRT image acquired in the next frame.

3.4 ROI detection

When studying the differences in the face caused by emotions, the main ROIs to analyse are forehead, nose or the whole face, [8], [9]. To this attempt we used Matlab Computer Vision System Toolbox, where it is possible to train a classifier based on the Viola-Jones algorithm for object detection and boosted cascade training [16]. Specifically, the function utilized for this purpose is CascadeObjectDetector. To deal with the non-static nature of our dataset, the classifier was trained with a balanced number of positive and negative ROIs. All the cases with a wrong ROI detection were manually corrected.

3.5 Efficiency Ratio

Due to the arithmetic nature of the TSST exercise performed (see Section 2), we defined a measure called Eficiency Ratio, as

$$ER = \frac{\text{Max. number of non stop right answers}}{\text{Total number of answers}} \quad . \quad (2)$$

For each single individual, ER is the portion that belongs to the longest series of correct answers over all the answered num-

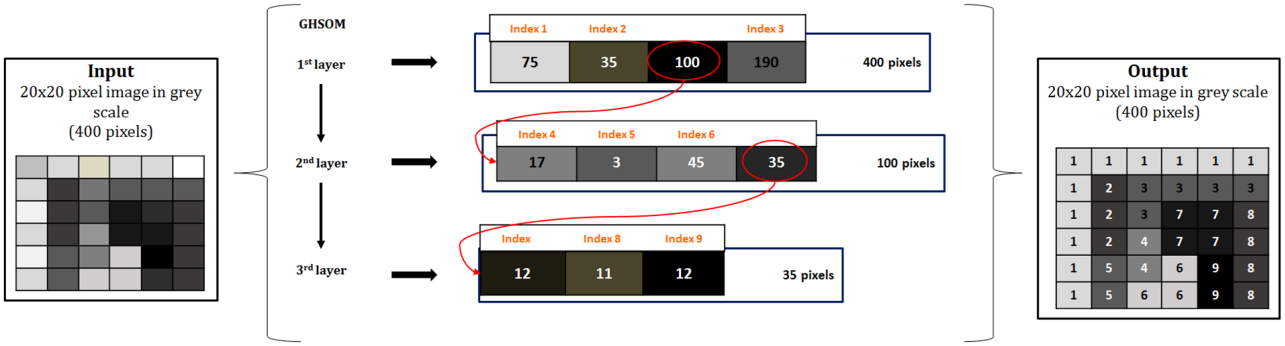


Figure 3. Example of an image classification by pixel value with Growing Hierarchical Self-Organizing Maps (GHSOM). The input consists of a 20x20 pixel grey image. In the first layer, 400 pixels are classified in four groups (four neurons). The group of 100 pixels represents more than δ portion of the total standard deviation so it is chosen to be the input for the second layer. Likewise, the group of 35 pixels is classified again as it has high standard deviation (measured by the parameter δ). The algorithm stops when the mean standard deviation of the groups in the layer is less than τ. Every group (neuron) has an index and its intensity value corresponds to the average of the pixel values in the group. The output image is the result of pixel classification.

bers. ER is used as a reference to measure the concentration of the subjects during the task.

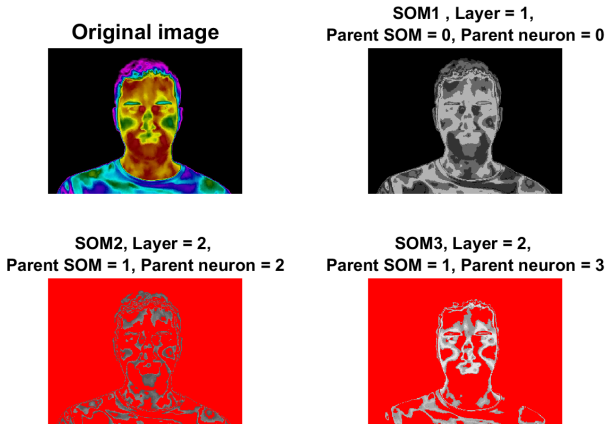


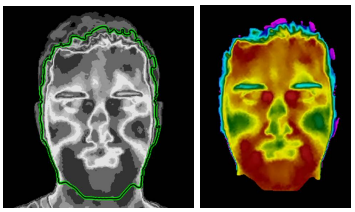
Figure 4. Infrared thermography image preprocessing with Growing Hierarchical Self-Organizing Maps of three layers with four neurons.

(a) (b)

Figure 5. (a) Output of Growing Hierarchical Self-Organizing Maps. (b) Absolute difference between the input and the output (image denoising).

4 Experimental results

All the images were preprocessed with a GHSOM of parameters $\tau = 0.3$ and $\delta = 0.03$, which results in a neural network of

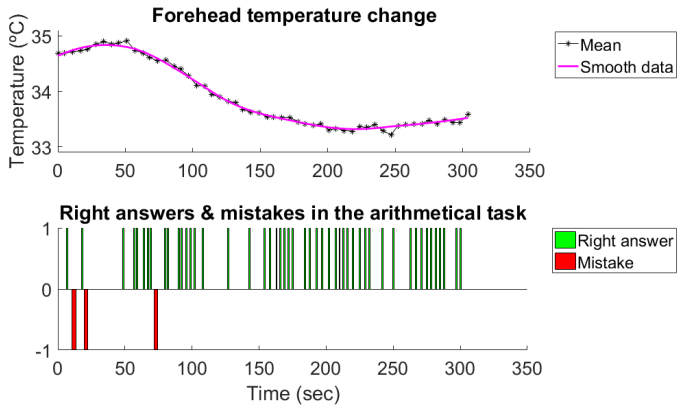


(a) (b)

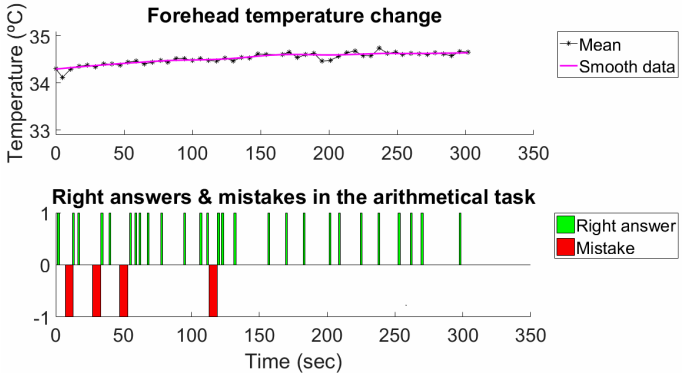
Figure 6. Result of the Hybrid Geodesic Region-Based Active Contours, (green curve). (a) Resulting active contour after Growing Hierarchical Self-Organizing Maps processing of the image. (b) Original image pixels in the obtained mask

three layers with four neurons each. Each layer represents specific temperature values (pixel intensities) (see Figure 4) and the noise is reduced in the output image (see Figure 5). The HGRBAC was set to a maximum of 160 iterations, a step of 0.2 to perform the gradient descent in the geodesic part of the eq. 1, and the radius of the neighborhood was set to 10% of the average length of the image. Figure 6 shows an example of the resulting segmentation using the HGRBAC. The CascadeObjectDetector for ROI detection was trained with a ground truth of 140 images (80 positive and 60 negative). The average error was 4%. The wrongly classified images were manually corrected in order to have all the information needed for the analysis.

The temperature of each ROI (i.e., face, forehead and nose) was measured as the average temperature of all the pixels that form the ROI, and it was calculated at each time point of the video -52 time points-. When calculating the temperature change from the beginning of the task to the end, two groups can be distinguished: a group of subjects who suffer an increase in temperature (5 subjects) and group who suffer a decrease (3 subjects). This classification applies to all ROIs. The Pearson correlation coefficient between the temperature of three ROIS shows us that the forehead and the face are strongly dependent ($\rho = 0.96$), and nose and forehead are weakly related ($\rho = 0.59$).



(a)



(b)

Figure 7. Graphical representation of forehead temperature change and task performance over time of two subjects. (a) (Top) Temperature at each time point (black dots) and an interpolation for the described function (pink). (Bottom) Representation of every right and wrong answer during the arithmetic task (right answers are the bars above the x-axis (green) and wrong answers are the bars under the x-axis (red)). (b) Same as (a).

The group that suffered a decrease in temperature over time shows a quadratic behaviour (Figure 7a) while the increase is linear (Figure 7b). In the examples of the Figure 7, the rise in temperature occurred while the subject made most of the mistakes. The subject with an increase in temperature got a smaller efficiency ratio ($14/31 = 0.45$) than the subject with a decrease ($42/53 = 0.79$).

The last result regarding the Efficiency Ratio is that two groups may be distinguished (see Figure 8): The group with values above 0.5 and the group with values below. The Efficiency Ratio range is $[0.35, 0.8]$, with an average value of 0.5278 and a standard deviation of 0.1908, which is quite high for a test that should be homogeneous for all the subjects.

5 Discussion

The presented method provides an accurate semi-automatic tool to analyze facial skin temperatures of single individuals recorded with an IRT camera. This approach supports the study of human emotions by a non-invasive method, which helps to set more realistic environments during the experiments. Unlike the techniques published so far, [8], [9], the proposed pipeline

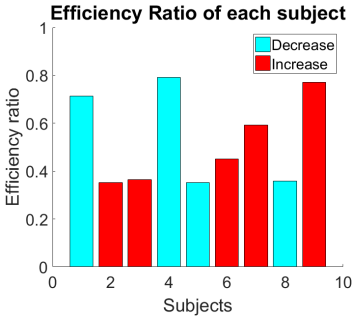


Figure 8. Bar plot of Efficiency Ratio by temperature behaviour. Only four subjects (44%) got an Efficiency Ratio higher than 0.5 in the arithmetic task, and five subjects (56%) performed under that scored. This classification is independent of the increase/decrease in temperature.

is able to compensate for the subject motion during the videos acquisition, resulting in a powerful tool for future research.

The use of active contours requires image denoising and manual initialization. We propose to utilize GHSOM hierarchical classification for image denoising. It has been shown that the proposed active contour converged to the correct boundary of the face. HGRBAC are robust to handle the characteristic non-homogeneous shapes of IRT images. Edge-based active contours were needed in order to evaluate the edges between the face, the ears and the hair, or the chin and the neck. As a future work, the initialization of the segmentation of the first frame should be included as a step of the algorithm to have a completely automatic method.

Temperature is represented by pixel intensities, so GHSOM classification leads to temperature values classification in the hierarchy, see Figure 4. It is possible then to choose the specific layer of the GHSOM to analyse the change of the desired temperature values over time and thereby, describe temperature patterns.

Due to the size of the dataset and the fact that the TSST was not entirely completed and repeated twice, patterns extracted from the analysis lack robustness and cannot be generalized. However, we encourage the reader to think about the two different patterns that appeared in this experiment and put them in relation to the results that have appeared in the literature up to this moment [8], [9]: Lineal and quadratic decreases in temperature. A feasible mathematical model for these behaviours has to cover both tendencies. The difference between both subjects is that one did not managed to concentrate, so he suffered emotional stress and the temperature of his forehead increased; and the second one, managed to concentrate most of the time, so he reduced the stress and suffered a decrease in temperature. This behaviour repeated in all the cases (linearity for the temperature increase and quadratic behaviour for temperature decrease). As far as we have seen in the literature, there is a large amount of work done to determine the temperature change caused by different affective statuses, but no studies have been done so far about the function described by the temperature.

We have found evidence about the effect of the concentration in human body thermal regulation. Recalling the studies of

Puri *et al*. [8] and Garbey *et al*. [7], it makes sense to think that the concentration affects the results and that it can be measured by bioheat modelling as long as we can measure the concentration. In fact, we tried to measure the latter by means of the Efficiency Ratio as it determines analytically how efficiently the individual is performing the exercise. Here, we assume that each subject has to be concentrated to perform the exercise successfully, and vice versa. Once again, two groups can be distinguished, but there is no evidence about the relationship with the temperature. So, we could only deduce that we managed to induce stress to one half of the subjects.

6 Conclusions

The pipeline presented in this work provides an accurate measure of the ROI temperature as captured in an IRT video. The algorithms used here are all available and the results can be reproduced so as to perform a more complete future study of human emotions.

Also, the results given by the stress analysis provide new information about the influence of subjects concentration and how the human body thermal regulation occurs in these cases. As it was discussed in previous lines, these results must be taken into account to design future studies aiming to characterize investigated cognitive tasks as performed by subjects.

7 Acknowledgements

We would like to acknowledge the support of the Spanish Ministry of Economy and Competitiveness, under grants number TEC2013-48552-C2-1-R (AMB, EGM), TEC2015-73064-EXP (AMB, EGM) and TEC2016-78052-R (AMB, EGM).
Thanks to Life Supporting Technologies Group (LST-UPM) for taking part on grant FIS-PI12/00514 (JF, APGM, AMUM).

References

[1] J. Aguiló, P. Ferrer-Salvans, A. García-Rozo, A. Armario, A. Corbí, F.J. Cambra *et al.*, "Project ES3: Attempting to quantify and measure the level of stress," *Revista de neurología*, vol. 61, no. 9, pp. 405–415, 2015.

[2] V. Engert, A. Merla, J. A. Grant, D. Cardone, A. Tusche, and T. Singer, "Exploring the use of thermal infrared imaging in human stress research," *PloS One*, vol. 9, no. 3, p. e90782, 2014.

[3] A. M. Guzman, M. Goryawala, J. Wang, A. Barreto, J. Andrian, N. Rishe, and M. Adjouadi, "Thermal imaging as a biometrics approach to facial signature authentication," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 214–222, 2013.

[4] J. F. Head, F. Wang, and R. L. Elliott, "Breast thermography is a noninvasive prognostic procedure that predicts tumor growth rate in breast cancer patients," *Annals of the New York Academy of Sciences*, vol. 698, no. 1, pp. 153–158, 1993.

[5] K. Ammer, "Diagnosis of Raynaud's phenomenon by thermography," *Skin Research and Technology*, vol. 2, no. 4, pp. 182–185, 1996.

[6] H. Keen, P. Mease, C. Bingham, J. Giles, G. Kaeley, and P. Conaghan, "Systematic review of MRI, ultrasound, and scintigraphy as outcome measures for structural pathology in interventional therapeutic studies of knee arthritis: focus on responsiveness," *The Journal of Rheumatology*, vol. 38, no. 1, pp. 142–154, 2011.

[7] M. Garbey, A. Merla, and I. Pavlidis, "Estimation of blood flow speed and vessel location from thermal video," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 1, pp. I–I, IEEE, 2004.

[8] C. Puri, L. Olson, I. Pavlidis, J. Levine, and J. Starren, "StressCam: Non-contact measurement of users' emotional states through thermal imaging," in *CHI'05 extended abstracts on Human Factors in Computing Systems*, pp. 1725–1728, ACM, 2005.

[9] S. Jenkins, R. Brown, and N. Rutterford, "Comparing thermographic, EEG, and subjective measures of affective experience during simulated product interactions," *International Journal of Design*, vol. 3, no. 2, 2009.

[10] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The Trier Social Stress Test – A tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[11] M. Dittenbach, D. Merkl, and A. Rauber, "The growing hierarchical self-organizing map," in *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 6, pp. 15–19, IEEE, 2000.

[12] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[13] S. Lankton, D. Nain, A. Yezzi, and A. Tannenbaum, "Hybrid geodesic region-based curve evolutions for image segmentation," in *Proc. SPIE Medical Imaging*, pp. 65104U–65104U, International Society for Optics and Photonics, 2007.

[14] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[15] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–I, IEEE, 2001.

Inter and Intra Class Correlation Analysis ($II_C CA$) for Human Action Recognition in Realistic Scenarios

S.Nazir*, M.H.Yousaf*, S.A.Velastin+

*University of Engineering and Technology Taxila, Pakistan
+Universidad Carlos III de Madrid, Spain

Keywords: Human action recognition, inter and intra class variation, correlation analysis, UCF Sports

Abstract

Human action recognition in realistic scenarios is an important yet challenging task. In this paper we propose a new method, Inter and Intra class correlation analysis ($II_C CA$), to handle inter and intra class variations observed in realistic scenarios. Our contribution includes learning a class specific visual representation that efficiently represents a particular action class and has a high discriminative power with respect to other action classes. We use statistical measures to extract visual words that are highly intra correlated and less inter correlated. We evaluated and compared our approach with state-of-the-art work using a realistic benchmark human action recognition dataset.

1 Introduction

Human action recognition is a commonly studied area in computer vision. Its expansion took off in the early 1980s [1]. A wide-ranging literature exists about action recognition in a number of fields, including computer vision, signal processing, machine learning, pattern recognition etc. Human action recognition has been studied for more than a decade and its importance has grown since, due to its applications in human safety and security, including video surveillance, human computer interaction, robot learning etc. Many authors have made efforts to review and classify different approaches as well as cite different useful applications [2, 3, 4, 5, 6].

Action recognition is a challenging task due to a substantial amount of variation in video data. Realistic environments contain challenges like cluttered background, scale and view point variation, occlusion, variation in subject appearance etc. [7]. Recognition performance decreases in complex and realistic environments [8].

In addition to these challenges, variation within different classes impact recognition performance. For instance, two different individuals walking will display differences in terms of stride length and speed. In addition, similarities between two different actions classes (e.g. jogging can be considered as walking at a higher speed) can lead to confusion. As shown in Fig.1 the difference between both actions is very small. In particular, variation in actions performed by different subjects with different gender and at different speed and style needs to be handled [9]. Actions that seem so different and contain well-defined gestures according to us, can vary when performed in an uncontrolled environment. Thus, a major challenge is to deal with the large variations in action classes. These inter and intra class variations have been reported in many papers [10, 11]. In this paper, the ultimate goal is to propose a generic system with higher discriminative power to have a clear separation amongst these variations.

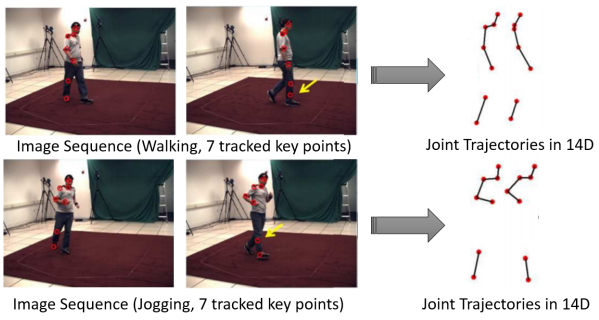


Figure 1. Walking and jogging action from the HumanEva videos dataset with their joint trajectories in 14D. (Courtesy [9]).

Wang et al. [12] proposed the actionlet ensemble method to handle the intra class variation in human actions. They captured human object interaction using LOP (Local Occupancy) features. Further, they represented each action as an actionlet: a conjunction of features for a subset of joints. They used a data mining approach to learn actionlets that are discriminative enough for each action class representation. They evaluated their approach on MSR-Action3D and MSRDaily Activity3D datasets. To handle inter and intra class variation Zunino [13] proposed a two-level classification approach. They recognize an action performed by a specific subject after identifying the subject first. They introduced a new evaluation strategy, known as personalization, to learn how actions are performed by a specific subject. Statistical measures were used to retrieve the subject specific role for handling inter and intra class variation challenge. They achieve reasonable performance on MSR-Action3D, MSRC-Kinect12 and HDM-05 benchmark datasets.

Here, we propose a new approach to analyze inter and intra class similarities between different action classes using correlation analysis. We aim to learn a model which can improve the

knowledge of our system by training it with class specific properties. We enrich our system with the information that is similar within a class and have a high discriminative power with respect to other action classes. During training, we select the adequate discriminative visual words representation for a particular action class and ignore the visual words with lower discriminative power using correlation analysis. As a result, our proposed approach provides significant results on a state-of-the-art realistic dataset for human action recognition.

2 II_CCA for Human Action Recognition

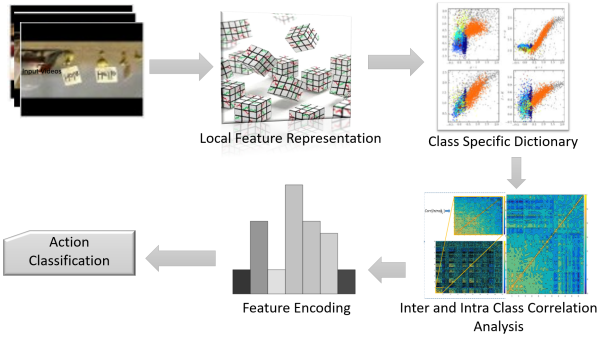


Figure 2. Proposed framework: II_CCA for human action recognition in realistic scenarios.

In this section, we discuss the proposed framework for realistic scenarios. Our approach follows a popular approach for human action recognition i.e. Bag of Words [14]. Our main focus is handling inter and intra class similarity challenges for human action recognition in a realistic environment. As shown in Fig.2 we represent each video using local feature representation, followed by the generation of class specific dictionaries for each action class. We use these class specific visual words representation for handling inter and intra class similarities using correlation analysis. After selecting highly intra correlated and less inter correlated visual words, we concatenate them into a visual word codebook. In the next step we encode video features using the visual words codebook and train a supervised classifier for action classification. The proposed approach is discussed in more detail in the following subsections.

2.1 Local feature representation

The first step in action recognition is to extract and represent features that are discriminative with respect to visual appearance and body movement of human body. To illustrate our approach, we use 3D Harris space time interest point detector as proposed by Laptev [15]. We use this detector to obtain interest points that are well localized in spatio-temporal domain and corresponds to meaningful events. These space time interest point corresponds to the non-constant motion in space time neighborhood. As proposed by Laptev [15], we have not adapted these interest point to scale and velocity so as to get sufficient interest points for video representation. These detected interest points are further described using a 3D SIFT descriptor [16]. 3D SIFT provides robustness to noise and orientation by encoding information in both space and time domain. It should be pointed out that the proposed method is independent of the feature representation chosen. Let $feat_i=\{f_1,f_2,f_3,....,f_p\}$ represent the total feature extracted for the i^{th} action class and p is the total number of features. These extracted features are grouped together as $FSet=\{feat_i\}^c{}_{i=1}$ where c is the total number of unique action classes.

2.2 Supervised class specific dictionary

The next step is to learn a dictionary that is discriminative enough to differentiate between different action class representations. Consider the feature representation of videos from each action class grouped together as $FSet=\{feat_i\}^c{}_{i=1}$ where c is the total number of unique action classes. We intend to learn a concatenated dictionary which consists of c class specific dictionaries $\phi_1, \phi_2, \phi_3,\ldots., \phi_c$.

Class specific dictionary learning has the advantage of class specific learning independently of other action classes [17]. It also have an advantage of parallel implementation as compared to classical dictionary learning technique [18]. Each class dictionary will have an *efficient* representation of its specific class and will be *less efficient* for other action classes.

2.3 Correlation analysis

2.3.1 Intra class correlation analysis

This section focuses on the correlation analysis of variation within each action class. Visual words from each action class ϕ_i are compared with themselves to obtain the relation between them. Let ϕ_i denote the visual words for the i^{th} action class. There are many ways to represent the correlation between different variables e.g. Pearson, Kendall and Spearman correlation. Pearson, Kendall and Spearman correlation coefficients are used to represent the association between two linear, ordinal and non-linear variables respectively. Because of the non-linear relationship, we measure the correlation between different visual words using the Spearman coefficient [19]. Spearman Correlation coefficient is computed as:

$$corr(\phi_i) = 1 - \frac{6d_w^2}{n(n-1)} \quad (1)$$

Where $d_w = r_g(\phi_i(w)) - rg(\phi_i(w))$ is the difference between two ranks of each visual word w and n is the total number of visual words in the i^{th} action class. The resultant matrix is denoted as $Corr(\phi_i)$for the i^{th} action class as shown in Fig.3(a). We consider this resultant matrix for mean correlation analysis. We used these statistics to characterize similarities and differences within a particular action class. Mean correlation is obtained by calculating the mean of all entries in $Corr(\phi_i)$ matrix and is represented as $MCorr(\phi_i)$. Mean correlation shows the degree of knowledge of each visual word within that class. The main diagonal elements of $Corr(\phi_i)$ shows the correlation of each visual word with itself, which is normally 1 as shown in Fig.3.

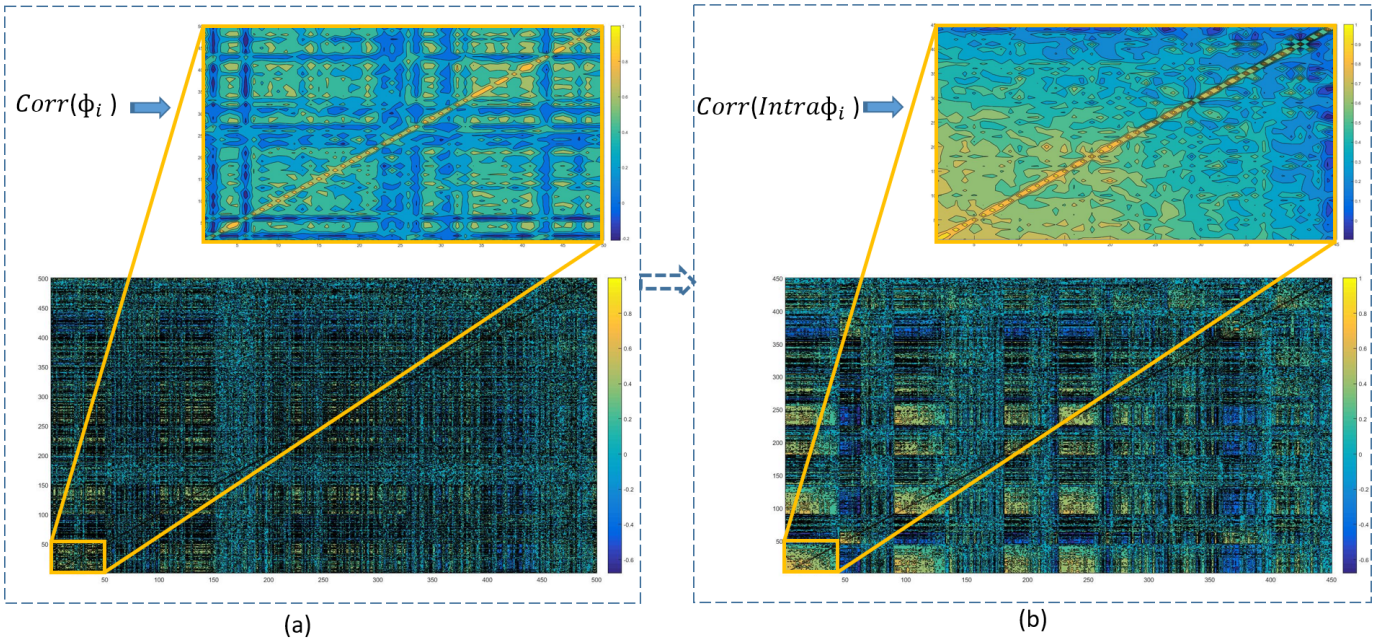


Figure 3. Correlation metrics for i^{th} action class (a) before intra class correlation analysis (b) after intra class correlation analysis.

The next step is the calculation of square root of mean correlation matrix. Square root of mean correlation shows the knowledge of each visual word among other visual words within its action class. Square root of mean correlation is denoted as $SMCorr(\phi_i)$ for i^{th} action class.

For maximizing intra class similarity, we have selected only those r visual words which are highly correlated with other visual words within its action class. In other words, we have only selected those visual words that efficiently represent its action class. Thus, the highly correlated visual words are denoted as $Intra\phi_i$ for the i^{th} action class. Correlation of these selected visual words is shown in Figure 3(b). High correlation is observed between the visual words representation of diving, kicking and riding horse action classes for UCF Sports dataset.

2.3.2 Inter class correlation analysis

In this section, we explain the measurement of variation between different action classes for inter class correlation analysis. We begin by computing the correlation between two different action classes. Let $Intra\phi_i$ and $Intra\phi_j$ represent the highly intra correlated visual words for the i^{th} and j^{th} action class respectively. We calculate the correlation between $Intra\phi_i$ and $Intra\phi_j$ using Spearman correlation coefficient computed as:

$$corr(Intra\phi_i, Intra\phi_j) = 1 - \frac{6d_w^2}{r(r-1)} \qquad (2)$$

Where $d_w = r_g(\phi_i(w)) - rg(\phi_j(w))$ is the difference between two ranks of each visual word w and r is the total number of highly intra correlated visual words in the i^{th} and j^{th} action class. The resultant matrix is denoted as Corr($Intra\phi_i$, $Intra\phi_j$). The resultant matrix consists of four quadrants, the upper right and lower left quadrant shows the correlation between $Intra\phi_i$ and $Intra\phi_j$. The upper left quadrant shows the correlation between $Intra\phi_i$ and the lower right quadrant shows the correlation between $Intra\phi_j$ as shown in Fig.4(a).

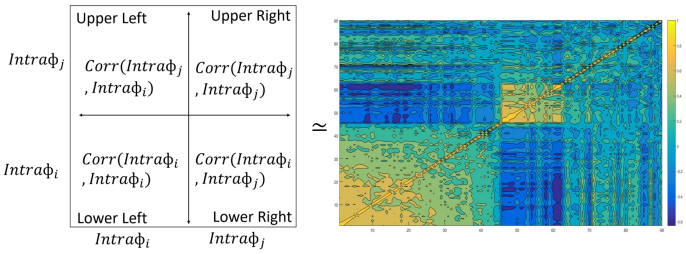


Figure 4. Correlation matrix for $Intra\phi_{ji}$ and $Intra\phi_j$.

As shown, using a graphical representation, in Fig.4(b) the upper right quadrant is the transpose of the lower left quadrant. We only choose the quadrant representing the correlation between $Intra\phi_i$ and $Intra\phi_j$ i.e. upper right quadrant or lower right quadrant for inter class correlation analysis. In the next step we calculate the mean correlation of these quadrants and denoted it as $MCorr(Intra\phi_i, Intra\phi_j)$. Followed by calculation of its square root, we represent it as $SMCorr(Intra\phi_i, Intra\phi_j)$.

Similarly to the concept discussed in the previous section, mean correlation matrix shows the degree of knowledge of each visual word of class i for action class j. The Square root of mean correlation shows the average knowledge of each visual word of i action class among visual words of j^{th} action class.

For minimizing the inter class similarity between different classes we have only selected those visual words of i^{th} action class which show less correlation with visual words of j^{th} action class. Here j varies from $1....c$ except i^{th} action class and c is the total number of unique action classes. In other words, we

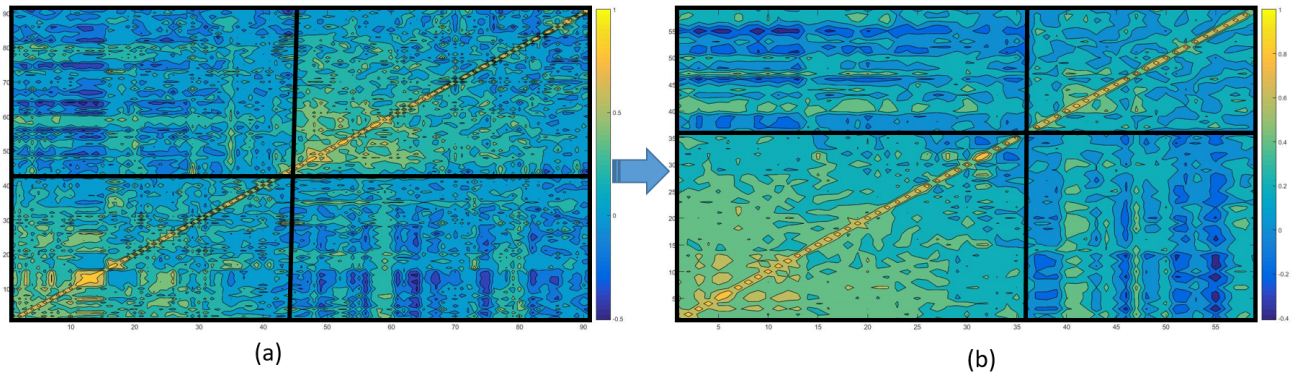


Figure 5. Correlation metrics the for i^{th} and j^{th} action classes (a) before inter class correlation analysis $Corr(Intra\phi_i, Intra\phi_j)$ (b) after inter class correlation analysis $Corr(Inter\phi_i, Inter\phi_j)$.

have only selected those visual words for the i^{th} action class that less efficiently represent the j^{th} action class. As a result, these selected visual words after inter class correlation analysis are represented as $Inter\phi_i$ for i[th] action class. Fig.5(b) shows the correlation of $Inter\phi_i$ and $Inter\phi_j$ class. The number of visual words for $Inter\phi_i$ are less than the number of visual words for $Intra\phi_i$ as shown in Fig.5(a).

2.3.3 Feature Encoding

After correlation analysis, we form a concatenated codebook $D=\{Inter\phi_1, Inter\phi_2, \ldots, Inter\phi_c\}$ where c is the total number of unique action classes. In this step, the main focus is encoding features for each video using the codebook D.

Let $F=\{f_1, f_2, f_3, \ldots, f_x\}$ represent the feature for each video. For each feature f_k the codebook word d_m can be viewed as a function of f and defined as

$$w(f) = \begin{cases} 1, & \arg\min_j ||f_k - d_j||_2 \\ 0, & otherwise \end{cases} \quad (3)$$

Where each feature vector votes for only its nearest codebook word. The occurrence of these votes are stored in histogram for each video.

2.4 Action Classification

Each video is represented as a histogram of highly intra class and less inter class correlated visual words. Further we train different classifiers using these video representation. We consider four different types of classification methods for training: Support Vector Machine (SVM), Nearest Neighbor Classifier (KNN), Decision Tree and Linear Discriminant analysis (LDA). For our experiments Linear discriminant analysis used empirical prior probabilities to determine class probabilities and KNN is trained by varying the number of nearest neighbors and $k=25$ provides the best result for as stated in Table.1. Decision tree considers $2^{c-1}-1$ combinations to predict the best split for class predictor where c is the total number of action classes. SVM used Gaussian kernel for learning which is defined as:

$$G(x_1, x_2) = exp(||x_1 - x_2||^2) \quad (4)$$

For all these classifiers we used the same cost measure which is defined as:

$$Cost(x, j) = \begin{cases} 1, & i \neq j \\ 0, & i == j \end{cases} \quad (5)$$

Our experimentation results shows that SVM performs better with respect to other three classifiers. SVM has also become a popular classifier for human action recognition. Our results show that Multiclass non-linear SVM trained using *c-1* binary support vector machine and the '*ordinal*' coding design scheme performs better, here c is the number of unique action classes.

3 Performance Evaluation

To test our approach, we performed number of experiments on publicly available dataset. All experiments were carried out on an Intel Core i7-6500U CPU with 2.50 Ghz, and the proposed algorithm was implemented in MATLAB 2015R(a).UCF Sports contains sports action videos captured in realistic environment. It contains 10 sports actions e.g. walking, diving, kicking, horseback riding etc. UCF Sports action videos have a large number of intra and inter class variation typical of many real life environments. We used leave one out cross validation method as proposed in [20].

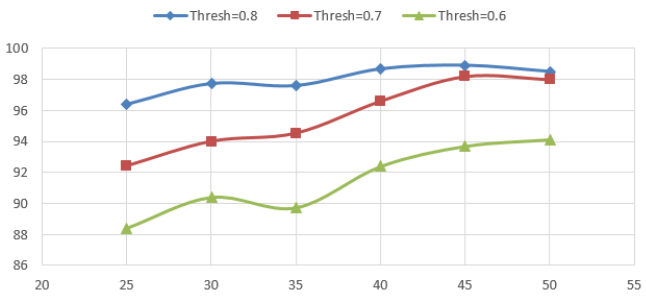


Figure 6. Parameter *(r and thresh)* evaluation for $II_C CA$.

For class specific dictionary construction we performed k-mean clustering using 50 visual words for each action class. In the next step, we performed different experiments to obtain

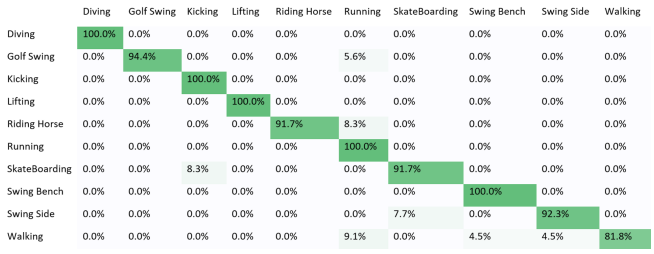


Figure 7. Confusion matrix for UCF Sports.

optimized *'r'* number of highly intra correlated visual words as show in Fig.6. Ignoring visual words representation for an action class by selecting smaller values of *r* can decrease performance. We further selected only those visual words that less efficiently represent other action class. We selected a threshold *'thresh'* by performing various experiments as shown in Fig.6. We have selected only those visual words that have less correlation with other action classes than this threshold.

Table 1. Classification method evaluation for II_CCA.

Classifier	**Accuracy**
Support Vector Machine	**98.90%**
K Nearest Neighbors	95.73%
Linear Discriminant Analysis	92.93%
Decision Tree	91.07%

Table 2. Comparison with state-of-the-art work for UCF Sports dataset.

Method	**Accuracy**
II_CCA (with SVM)	**98.90%**
CNN + Rank Pooling [21]	87.20%
Dense Trajectories + MBH [22]	88.00%
Independent sub space analysis [23]	86.50%

In our last experiment, we evaluated our approach using different classification methods for r=45 and thresh=0.8. As shown in Table.1, SVM proves to perform better with respect to other classifiers. Fig.7 shows the resultant confusion matrix for the UCF Sports dataset when evaluated using selected parameter as described above. As expected, significantly less inter and intra class similarity is observed between different classes for a realistic dataset (i.e. UCF Sports). Table.2 shows comparison with some other methods for human action recognition for UCF Sports dataset. Our approach achieves better performance as compared to other state-of-the-art methods.

4 Conclusion

In this paper, we have proposed a new approach to handle inter and intra class variation in realistic scenarios. We have shown that by computing the correlation between visual representations for each action class, we can handle inter and intra class variation challenge. First we learn class specific visual representation for each action class. Further we exploit these visual representations that have high intra class similarity and low inter class similarity. Finally, we demonstrate the potential of our proposed Inter and Intra class correlation analysis (II_CCA) approach for action recognition by evaluating its performance on a realistic human action recognition dataset. Future work will be to strengthen the robustness of our approach to other challenges like occlusion and view invariance for human action recognition in realistic scenarios.

Acknowledgements

S.A. Velastin has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nº 600371, the Ministerio de Economía, Industria y Competitividad (COFUND2013-51509) the Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander.

References

[1] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 20–27, IEEE, 2012.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[4] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[5] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[6] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, pp. 568–576, 2014.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[9] D. Gong and G. Medioni, "Dynamic manifold warping for view invariant action recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 571–578, IEEE, 2011.

[10] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[11] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 588–595, 2014.

[12] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1290–1297, IEEE, 2012.

[13] A. Zunino, J. Cavazza, and V. Murino, "Revisiting human action recognition: Personalization vs. generalization," *arXiv preprint arXiv:1605.00392*, 2016.

[14] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[15] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[16] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 357–360, ACM, 2007.

[17] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.

[18] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognition*, vol. 45, no. 11, pp. 3902–3911, 2012.

[19] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficient on the same sets of data," *Quaestiones Geographicae*, vol. 2, no. 30, 2011.

[20] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[21] B. Fernando and S. Gould, "Learning end-to-end video classification with rank-pooling," in *Proc. of the International Conference on Machine Learning (ICML)*, 2016.

[22] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[23] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3361–3368, IEEE, 2011.