Affect Analysis and Membership Recognition in Group Settings

Wenxuan Mou

School of Electronic Engineering and Computer Science

Queen Mary, University of London

This dissertation is submitted for the degree of

Doctor of Philosophy

Multimedia and Vision Group

March 2020

I would like to dedicate this thesis to my loving parents and beloved husband.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Wenxuan Mou

March 2020

Acknowledgements

I would like to express my sincere gratitude to all the people who supported me during my PhD. The work presented in this thesis would not have been possible without the guidance of my supervisors, Dr. Hatice Gunes and Prof. Ioannis Patras. I thank them for their support, encouragement and patient guidance throughout the PhD project. I am thankful to my independent accessors, Dr. Timothy Hospedales and Dr. Matthew Purver, for many useful comments. My thanks also go to my tutor and collaborators at MERL, Dr. Tim Marks, Dr. Chen Feng and Dr. Xiaoming Liu, for providing me an opportunity to work as an intern for 6 months in a very friendly research environment.

I would like to thank my family specially, my husband, my parents and my sister, who have been supporting, encouraging, accompanying and helping me all the time during my PhD. I knew I can always rely on them when I have any difficulties.

My warmest thanks go to my friends and colleagues at Queen Mary University of London, Ye Tao, Zhaoyang Xu, Tingting Xie, Dr. Shenglang Huang, Dr. Petar Palasek, Dr. Christos Tzelepis, Aria Ahmadi, Mina Bishay, Dr. Youngkyoon Jang, Dr. Juan Abdón Juan Abdón Miranda Correa, Dr. Ioannis Marras, Silvia Cascianelli, Dr. Heng Yang, Dr. Oya Celiktutan, Dr. Faranak Sobhani. Especially I would like to thank Dr. Juan Abdón Juan Abdon Miranda Correa for collecting and annotating the databases, Mina Bishay for annotating the databases, Dr. Christos Tzelepis, Dr. Heng Yang, and Dr. Oya Celiktutan for the collaboration.

I really appreciate my funding body China Scholarship Council and Queen Mary University of London for the support of my PhD studies. I would also like to thank the British

Machine Vision Association, the Henry Lester Trust and the Great Britain-China Educational Trust for their generous financial support during my PhD project.

Abstract

Emotions play an important role in our day-to-day life in various ways, including, but not limited to, how we humans communicate and behave. Machines can interact with humans more naturally and intelligently if they are able to recognise and understand humans' emotions and express their own emotions. To achieve this goal, in the past two decades, researchers have been paying a lot of attention to the analysis of affective states, which has been studied extensively across various fields, such as neuroscience, psychology, cognitive science, and computer science. Most of the existing works focus on affect analysis in individual settings, where there is one person in an image or in a video. However, in the real world, people are very often with others, or interact in group settings. In this thesis, we will focus on affect analysis in group settings. Affect analysis in group settings is different from that in individual settings and provides more challenges due to dynamic interactions between the group members, various occlusions among people in the scene, and the complex context, e.g., who people are with, where people are staying and the mutual influences among people in the group. Because of these challenges, there are still a number of open issues that need further investigation in order to advance the state of the art, and explore the methodologies for affect analysis in group settings. These open topics include but are not limited to (1) is it possible to transfer the methods used for the affect recognition of a person in individual settings to the affect recognition of each individual in group settings? (2) is it possible to recognise the affect of one individual using the expressed behaviours of another member in the same group (i.e., cross-subject affect recognition)? (3) can non-verbal behaviours be used for the recognition of contextual information in group settings?

In this thesis, we investigate the affect analysis in group settings and propose methods to explore the aforementioned research questions step by step. Firstly, we propose a method for individual affect recognition in both individual and group videos, which is also used for social context prediction, i.e., whether a person is alone or within a group. Secondly, we introduce a novel framework for cross-subject affect analysis in group videos. Specifically, we analyse the correlation of the affect among group members and investigate the automatic recognition of the affect of one subject using the behaviours expressed by another subject in the same group or in a different group. Furthermore, we propose methods for contextual information prediction in group settings, i.e., group membership recognition - to recognise which group of the person belongs. Comprehensive experiments are conducted using two datasets that one contains individual videos and one contains group videos. The experimental results show that (1) the methods used for affect recognition of a person in individual settings can be transferred to group settings; (2) the affect of one subject in a group can be better predicted using the expressive behaviours of another subject within the same group than using that of a subject from a different group; and (3) contextual information (i.e., whether a person is staying alone or within a group, and group membership) can be predicted successfully using non-verbal behaviours.

Contents

List of Figures

List of Tables

Nomenclature

BoW	Bag-of-Words
AUs	Action Units
CCC	Concordance Correlation Coefficient
CNNs	Convolutional Neural Networks
ECG	Electrocardiogram
EEG	Electroencephalogram
ELEA	Emergent LEAdership
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
LBP	Local Binary Patterns
LBP-TOP	Local Binary Pattern histograms from Three Orthogonal Planes
LPQ	Local Phase Quantization
LPQ-TOP	Local Phase Quantization histograms from Three Orthogonal Planes
MAE	Mean Absolute Error

MSDF	Multi-Scale Dense SIFT
MSE	Mean Squared Error
PCC	Pearson Correlation Coefficient
PHOG	Pyramids of Histograms Of Gradients
QLZM	Quantized Local Zernike Moment
QoVE	Quality of the Viewing Experience
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
*k*NN	*k*Nearest Neighbours
FV	Fisher Vectors
GMM	Gaussian Mixture Model
HOF	Histograms of Optical Flow
HOG	Histogram of Oriented Gradients
PCA	Principal Component Analysis
SGD	Stochastic Gradient Descent

Chapter 1

INTRODUCTION

1.1 Motivation of the thesis

> *"Human behaviour flows from three main sources: desire, emotion, and knowledge."*
>
> — Plato

Emotions play an important role for humans in how we think, behave and communicate. The emotions we feel every day can compel us to take action, influence the decisions we make about our lives, both large and small, as well as can help us communicate with others more effectively. Interests into human emotions can be dated back to the Golden Age of Pericles' Athens, when the philosophical analysis of emotions was introduced by Plato and developed further by Aristotle (Knuuttila, 2018).

After computers came to our daily life decades ago, in order to advance human-computer and human-robot interaction, it is important to enable machines to understand and express emotions. Even though with the development of artificial intelligence and more intelligent machines are appearing in our daily life, only when these machines can understand feelings of humans and express their emotions, can they communicate with humans effectively and blend into our day-to-day life. To address this challenge, affective computing, proposed by Professor Rosalind Picard in her seminal paper (Picard, 1995) in 1995, has developed into an interdisciplinary field spanning computer science, psychology, and cognitive science. It aims to study and develop systems and devices that can recognise, interpret, process and simulate human feelings and emotions.

Recognition and analysis of human affects have been attracting increasing interest in the past two decades and have been applied in very diverse areas (Dautenhahn, 2007; Hernandez et al., 2012; Kleinsmith and Bianchi-Berthouze, 2013). In addition to the human-computer interaction (Barros et al., 2015) mentioned above, in assistant driving, vehicle driver's attention/engagement level can be detected through emotion analysis (Cai and Lin, 2011); in the security field, suspicious behaviours can be detected and tracked by analysing human emotions in surveillance videos (Arunnehru and Geetha, 2017); in healthcare, pain (Sikka et al., 2013; Bartlett et al., 2014) and depression (Joshi et al., 2012; Jain et al., 2014) can be detected for monitoring vulnerable people; in education, engagement can be monitored to track the attention of the student especially for e-learning (Niu et al., 2018; Yang, Wang, Peng and Qiao, 2018).

As highlighted above, affective computing is a very active research field that has received a lot of attention and achieved substantial progress over the last two decades. However, in order to advance the state of the art, and make affective computing research applicable to everyday applications, there are still a number of open issues that need further investigation. The majority of existing works focus on affect analysis in individual settings where there is only a single person in the image or video. In contrast, in the real world people are often being with others and interacting with each other in a group in their daily social life. Social psychologists have found that human behaviours are largely dependent on social context, i.e., the way humans behave alone is different from how humans behave in a group setting where two or more than two people appear in the scene (Barsade and Gibson, 2012).

However, little attention has been paid to affect analysis in group settings, either at the overall group-level emotion displayed by the whole group collectively or at the individual-level emotion displayed by each individual within a group.

Affect recognition in group settings has great potential to be used in various applications in the real world. For instance, in education, to assist teachers in obtaining a better insight of the students and how the learning taking place, the analysis of the emotions and engagement of each student in the class and the whole class is needed. In marketing, when analysing the expressions of customers, there are usually multiple customers in a scene (McColl-Kennedy et al., 2009). In human-robot interaction, to advance the collaboration and communication between robots and humans, robots need to identify humans' emotion in complex environments in which multiple people exist (Dautenhahn, 2007). In sports and entertainment events, to profile the mood of the audience in concerts and Olympics, the overall group-level or collective emotion displayed by the audience can be important for telecast audience ratings of such events (Sintsovaa and Musata, 2013).

As mentioned before, most of the existing methods for automatic emotion recognition focus on the analysis in individual settings, that is when a single person is in an image or in a video. One could consider transferring the models developed in such literature to solve the problems of automatic affect analysis in group settings directly [1]. However, there are many differences between the individual and group settings that make such attempts challenging. Compared to individual settings, group settings are more complex due to the group dynamics that are difficult to capture and can change (Lehmann-Willenbrock et al., 2017). The challenges on affect analysis in group settings include but not limited to these listed below.

1. Due to the differences between individual and group settings, individuals in group settings may express behaviours differently from being in individual settings, so that it

[1] Note that the group setting in this thesis refers to the audience setting with four people sitting together watching movies. A group refers to the group formed by the four people who are watching movies together.

is interesting to investigate whether the method used for affect recognition in individual settings can be transferred to group settings.

2. Context information in group settings is complex, but may have an effect on or even be used for affect recognition. The affect of an individual in group settings may not only be determined by what he/she is doing or where he/she is, but also what the other people in the group are doing and feeling (Barsade, 2002).

3. Context information is important but difficult to recognise, e.g., who individuals are being with, whether an individual is alone or in a group, what task people are engaged in, and acquaintanceship of the group members.

Not all challenges can be studied in this thesis, but we focus on the following research questions:

1. Is it possible to transfer the method used for the affect recognition of individuals in individual settings to group settings?
   This research question aims to investigate the aspects of Challenge 1.

2. Is it possible to recognise the affect of one individual using the expressed behaviours of another member in the same group?
   This research question aims to investigate the aspects of Challenge 2.

3. Can non-verbal behaviours be used for the context recognition (i.e., (1) whether an individual is alone or in a group, and (2) group membership of an individual - which group the individual is in)?
   This research question aims to investigate the aspects of Challenge 3.

To address the aforementioned research questions, in this thesis we concentrate on creating efficient approaches for affect analysis in group settings. The main contributions of this thesis are listed as follows.

1. Firstly, a framework is introduced to investigate whether the method used for affect recognition of a person in individual settings can be transferred to group settings, by utilising and systematically comparing different face and body features (Mou et al., 2019*a*).

2. Secondly, a novel framework is proposed to investigate whether it is possible to recognise the affect of one individual using the non-verbal behaviours of another group member. We first analyse the correlation of the affect among group members and then investigate the automatic recognition of the affect of one subject using the expressive behaviours of another subject in the same group (Mou et al., 2019*b*).

3. Thirdly, methodologies are presented for contextual information prediction using non-verbal behaviours, i.e., (1) whether a person is being alone or within a group and (2) if an individual is within a group, which group the individual belongs to (group membership recognition) (Mou et al., 2019*a*; 2018).

With the above investigations, the conclusions are summarised below.

1. The experiments show that the methods used for affect recognition of a person in individual settings can be transferred to group settings. Both face and body behaviours that are commonly used for affect recognition in individual settings are also shown efficient in affect recognition in group settings.
   This is corresponding to the contribution 1.

2. The experiments show that (1) the affect of people in the same group do correlate more than that of people in different groups; and (2) the affect of one subject in a group can be better predicted using the expressive behaviours of another subject within the same group than using that of a subject from a different group. These results are different from the findings of Hess, Banse and Kappas (Hess et al., 1995). Firstly, (Hess et al., 1995) did not perform automatic affect recognition, but in this thesis we do recognise

the affect automatically along arousal and valence dimensions. Secondly, in (Hess et al., 1995), the intensity of emotions is compared between a person being alone and being with another person; in this thesis the correlation of the affect is compared between people in the same group and people in different groups.

This is corresponding to the contribution 2.

3. A set of experiments show that it is possible to predict the context information using non-verbal behaviours, i.e., (1) whether an individual is alone or in a group and (2) group membership of an individual who is in group settings.

   This is corresponding to the contribution 3.

1.2 List of publications

The works presented and discussed in this thesis has resulted in the following peer reviewed publications:

Journal articles

- **W. Mou**, H. Gunes and I. Patras, "Alone vs In-a-group: A Multi-modal Framework for Automatic Affect Recognition.", submitted to *ACM Transactions on Multimedia Computing, Communications, and Applications*.

- **W. Mou**, C. Tzelepis, H. Gunes, V. Mezaris and I. Patras, "A Deep Generic to Specific Recognition Model for Group Membership Analysis using Non-verbal Cues." *Image and Vision Computing*, 2018.

Conference & workshop papers

- **W. Mou**, H. Gunes and I. Patras, "Your Fellows Matter: Affect Analysis across Subjects in Group Videos." submitted to Proceedings of *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2019.

- **W. Mou**, C. Tzelepis, H. Gunes, V. Mezaris and I. Patras, "Generic to Specific Recognition Models for Membership Analysis in Group Videos." Proceedings of *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2017.

- **W. Mou**, H. Gunes and I. Patras, "Alone versus In-a-group: A Comparative Analysis of Facial Affect Recognition." Proceedings of *ACM Multimedia Conference (ACMMM)*, 2016.

- **W. Mou**, H. Gunes and I. Patras, "Automatic Recognition of Emotions and Membership in Group Videos." Proceedings of *International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2016.

1.3 Outline of the thesis

The thesis consists of seven chapters and an overview of each chapter of the thesis is shown as follows:

- Chapter 1 introduces the motivation of the research, lists the main contributions, and outlines the thesis.

- Chapter 2 presents the commonly used definitions and models in affective computing, introduces the most widely used datasets in affective computing and presents the details of the AMIGOS dataset that is used in this thesis.

- Chapter 3 introduces the rapid development of affective computing; describes the commonly methods in both individual and group settings; and reviews the existing works in automatic analysis of group dynamics and other social dimensions in group settings, e.g., group cohesion and engagement.

- Chapter 4 presents a framework for affect analysis in both individual and group videos. It describes the extraction of face and body features; and reports the experimental results of affect recognition along both arousal and valence dimensions.

- Chapter 5 presents a novel framework for affect analysis across subjects in group videos. It analyses the correlation of the affect among group members and presents affect recognition results of one subject using the behaviours expressed by another subject in the same group.

- Chapter 6 presents a novel framework for group membership recognition. That is to recognise which group an individual belongs to using body behaviours. Extensive experimental results are provided and discussed.

- Chapter 7 concludes the thesis; discussions together with recommendations for future research and practical applications are also provided in this chapter.

Chapter 2

EMOTION THEORY & DATABASES

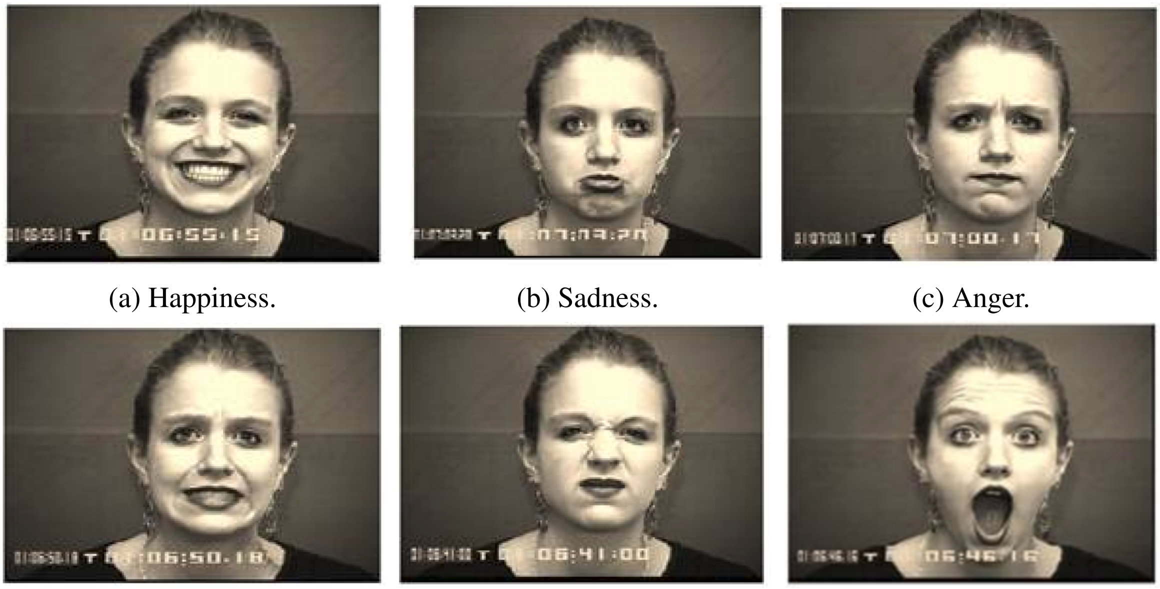
2.1 Definitions of emotion

As we aim to detect and recognize affect automatically, it is important to define how to represent affect in affective computing. Note that in this thesis, we use the terms emotion, affect and affective state interchangeably. We review the two approaches to define and model emotions, i.e., categorical approach and dimensional approach [1] (Gunes and Schuller, 2013).

2.1.1 Categorical approach to represent affect

Categorical approaches are the most commonly used approaches for computational analysis of affect. Categorical approaches enable representation of affective states and emotions using a predefined set of categories (or classes) such as neutral, fear, happiness, sadness, surprise, anger, and disgust. In the current literature the most widely used emotion classes are the basic emotions defined by Ekman and his colleagues (Ekman and Friesen, 2003). These basic emotions are modelled with six classes, namely, happiness, fear, surprise, sadness, anger and disgust as shown in Figure 2.1. To date, basic emotion categories proposed by Ekman have

[1] Both the categorical and dimensional approach here refer to the methods used for decoding the emotions from the expressive behaviours, such as facial behaviours, body gestures and vocal information (Mendolia, 2007; Scherer et al., 1991).



(a) Happiness.

(b) Sadness.

(c) Anger.

(d) Fear.

(e) Disgust.

(f) Surprise.

Figure 2.1 An illustrative figure to show the six basic emotions, i.e., happiness, sadness, anger, fear, disgust and surprise (Lucey et al., 2010).

been the most commonly adopted approach in research on automatic affect recognition (Tian et al., 2001; Mollahosseini et al., 2017; Dhall et al., 2012). However, in our real life, there are a number of non-basic, non-typical but more subtle and complex emotions, such as nervous, confused and excited. It is even found that there are more such emotions in daily life than so-called basic six emotions (Junek, 2007). Therefore, it is obvious that it is far from enough to represent emotions in our day to day life with these six basic ones. Under this situation, more recently researchers considered some alternative ways to model non-basic emotions. One approach is to add a limited number of emotion classes, such as relief and contempt, in addition to the six basic emotions (Bänziger et al., 2012).

2.1.2 Dimensional approach to represent affect

Dimensional approach refers to affect representation in a multi-dimensional space rather than using discrete labels like in the categorical approach. By employing a dimensional space, the dimensional approach can represent a wider range of emotions and continuously model the affect dimensions. Compared to the categorical approach, the dimensional approach is easier to describe more subtle and complex emotions that are difficult to describe by using a limited number of discrete emotions because of that emotions can be expressed in one or two or more continuous scales. For this practical reason, there are an increasing number of researchers paying attention to defining emotions based on dimensional approaches. Russell (Russell, 1980) defined the circumplex model to represent emotions into a 2D continuous space, i.e., <valence, arousal>. As shown in Figure 2.2, the dimensional approach is used to represent emotions in a 2D space, i.e., <valence, arousal>, and the discrete emotions are labelled in the 2D space. Here, valence is used to express how positive or negative people's emotions are. In this manner, for instance, "happiness" is one type of emotions in the "positive valence" region and "anger" is one type of emotions in the "negative valence" region, as shown in Figure 2.2. Arousal is used to express how active people are, being various from very calm to very excited. As can be observed in Figure 2.2, sleepiness is placed in the low arousal space whereas surprise and alarm appear on the other end of the arousal space.

One dimension and 3D space are also commonly used to represent affect. For example, only the dimension of valence is used to describe the emotion from negative to positive in (Sneddon et al., 2011). 3D space, <arousal, valence, dominance>, is often used (Poria et al., 2017). Here dominance is used to express how people's emotions are under control and it can be different from being overwhelmed to totally in control. The three dimensions can also be <evaluation, activation, power> and <pleasure, arousal, dominance> in the literature (Poria et al., 2017). In addition to categorical and dimensional approaches, Facial Action Coding System (FACS) model (Friesen and Ekman, 1978) is also widely used. FACS is a system describing all visually discernible facial movement and breaks down facial expressions into individual components of muscle movement, i.e., Action Units (AUs) (Friesen and Ekman, 1978). Based on FACS, a facial expression can be decomposed into one or more AUs. As

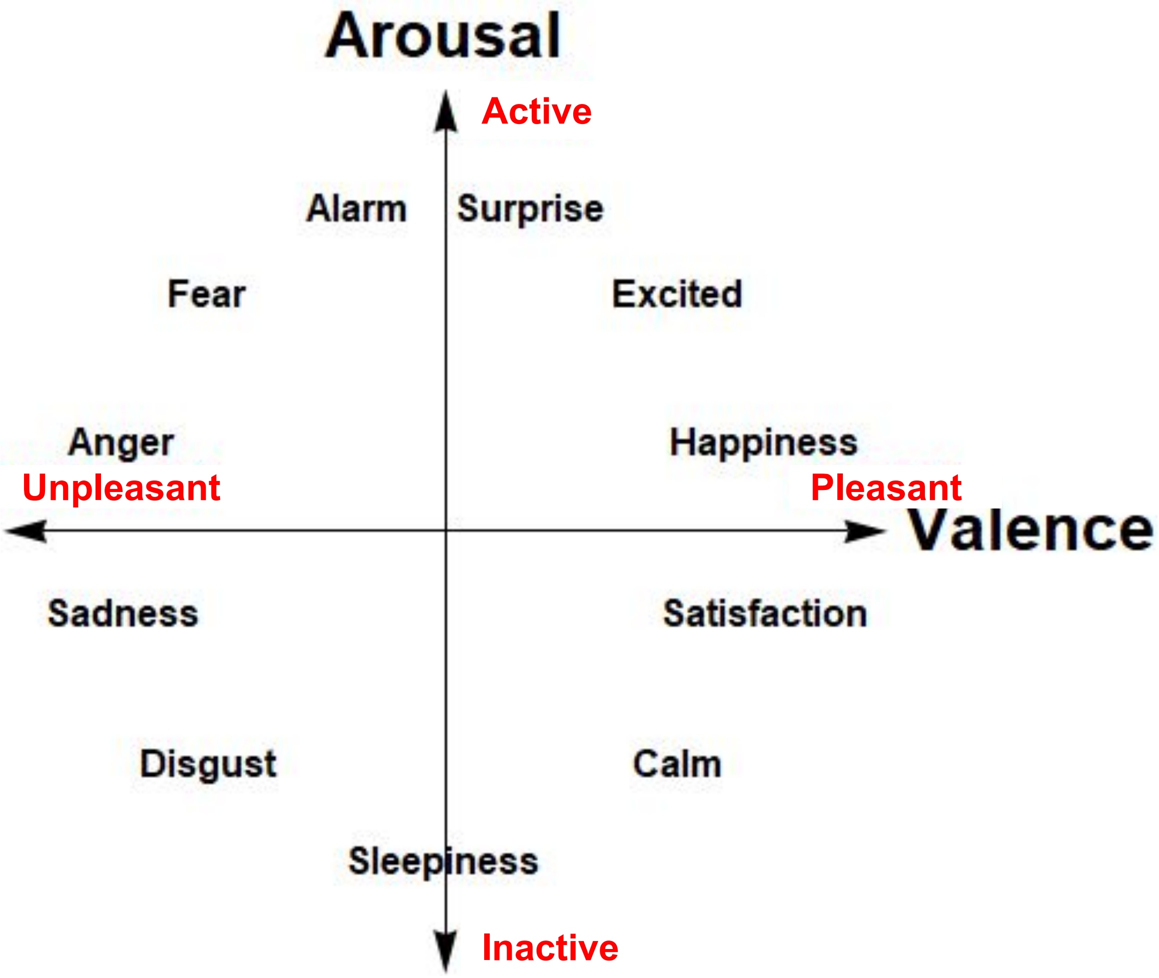


Figure 2.2 An illustration of 2D space to represent emotions, i.e., <valence, arousal> (Russell, 1980). The discrete emotions are also labelled in the space (Correa, 2018).

FACS or AUs are not the focus of this thesis, we will not discuss in details, but there are many works on facial AUs (Lucey et al., 2010; Valstar and Pantic, 2006).

2.2 Definitions of a group

A group. A collection of people is called a group (Forsyth, 2018). There is no doubt that a group always consists more than one person, however, there have been debates about whether dyads that consist two people should be called as a group (Moreland, 2010; Williams, 2010). It is argued that dyads have properties that are different from typical group process and some certain social phenomena cannot be studied in dyads, such as forming coalitions and any phenomena that require the study of subgroups (Moreland, 2010; Lehmann-Willenbrock et al., 2017). In addition, it is claimed that studying dyads is easier than studying larger groups that consist three or more people, but the results obtained from studying dyads cannot directly be used for groups containing three or more people (Lehmann-Willenbrock et al., 2017). In contrast, it is argued by some other researchers that the smallest group consists of two people (Williams, 2010; Chung et al., 2018). Williams (Williams, 2010) believes that 'the list of instances in which dyads are groups far exceeds the occasions when they are not'. One important argument is that the two fundamental aspects of group behaviours, i.e., social facilitation [2] and social loafing[3], can occur and can be studied in dyads. For more details, the readers are referred to (Moreland, 2010; Williams, 2010; Forsyth, 2018).

Small group. The size of a group varies a lot, from a few people to huge crowds, mobs, and assemblies (Forsyth, 2018). The size of a group has an effect on the other features or dynamics of a group. For example, people in a smaller group are linked with a strong emotional bond, while people in a larger group are rarely directly connected to all other members (Forsyth, 2018). There are many works focusing on the study of small groups (Sapru and Bourlard, 2015; Aran et al., 2010; Kelly and Barsade, 2001; Sanchez-Cortes et al., 2012; Hung and Gatica-Perez, 2010; Avci and Aran, 2014; Pai et al., 2015; Gatica-Perez, 2009). The number of people to form a small group varies in the literature. For example, a small group is defined as the group that consists of three to six people in (Gatica-Perez, 2009), while it is mentioned a small group usually consists of two to five people in (Pai et al., 2015). In this thesis, we focus on the study of small groups of four people.

Connections of group members. In a group, individuals are connected by and within social relationships (Forsyth, 2018). Every individual of the group does not need to be

[2] Social facilitation refers to that we behave differently in groups than we do when alone, and it tells us how we behave differently (Williams, 2010).

[3] Social loafing is defined as a reduction of individual effort when combining one's input with others (Williams, 2010).

always directly linked to every other individual in the group (Forsyth, 2018). There are different relations that link the individuals to be a group. For example, in families, the relationships are based on kinship, while in a football team, they are based on the task-related inter-dependencies[4]. Group members usually need to interact with each other to accomplish one or more goals. These interactions can be task-related or relational (Lehmann-Willenbrock et al., 2017; Forsyth, 2018). For task-related interaction, in most of the cases, group members coordinate their skills, resources and motivations to make a decision, produce a result, or make a product etc. The other interaction is the relational interaction, where creating and sustaining relationships is an outcome of groups (Lehmann-Willenbrock et al., 2017), such as providing support and suggestions to group members that need help (Forsyth, 2018). Both task and relational outcomes are accomplished by group members working interdependently, i.e., the group members depend on one another to get the outcomes of the group and the actions, thoughts and feelings of one group member are influenced by others in the group (Lehmann-Willenbrock et al., 2017; Forsyth, 2018).

Group dynamics. When anthropology, psychology, sociology, and the other social sciences emerged as unique disciplines in the late 1800s, the dynamics of groups became a topic of critical concern for all of them. Group dynamics refers to the behaviours and psychological processes that occur in a group or between groups over time (Lehmann-Willenbrock et al., 2017; Forsyth, 2018). For instance, groups tend to become more cohesive over time (Forsyth, 2018); larger groups may break down into smaller groups (Forsyth, 2018); the emotions of group members tend to converge with others (Barsade, 2002). Nowadays, with the development of the computational methods, researchers in computer science have been working on automatic analysis of group dynamics (especially for small groups) (Gatica-Perez, 2009), such as group cohesion (Hung and Gatica-Perez, 2010) and dominance in small groups (Aran et al., 2010; Hung and Gatica-Perez, 2010).

[4] The inter-dependence means that group members depend on one another; their outcomes, actions, thoughts, feelings and experiences are partially determined by others in the group (Forsyth, 2018).

2.3 Evaluation of affect analysis models

Affect recognition results obtained from different recognition systems are not always comparable against each other due to the fact that the implementation details of different works are quite different in terms of datasets (posed or non-posed data), the validation procedure, labels (e.g., discrete basic emotions, discrete non-basic emotions and continuous emotions) and evaluation metrics (e.g., mean squared error or Pearson Correlation Coefficient). However, there are still a number of commonly used evaluation methods for affect analysis, and we will review these methods in this section.

In order to train a generic model that can be used for an unseen subject, it is necessary to avoid the subject-dependence problem. In this case, a widely adopted validation procedure is leave-one-subject-out cross validation, which refers to using the data of one subject as validation set and data of all the other subjects as training data. Lucey *et al.* (Lucey et al., 2010) used leave-one-subject-out cross validation for both facial action unit and emotion recognition. The leave-one-subject-out cross validation was also used in (Bartlett et al., 2003) for the recognition of six basic emotions plus neutral. At a further step, cross-dataset validation, which trains the recognition model on one dataset and tests the learned model on another dataset, evaluates the generalisation ability of the recognition model (Sariyanidi et al., 2015). In addition to subject-independent validation, some works also adopted cross validation of affect analysis within each subject, i.e., subject-specific cross validation. In this case, the commonly used method is to do leave-one-sample-out cross validation for each subject (Abadi et al., 2015; Correa et al., 2018). To some extent, affect is subject specific, therefore, subject-specific cross validation can generate a better model for the target subject than using subject-independent cross validation method, i.e., leave-one-subject-out cross validation. Thus, in the case that an affect recognition model for a certain subject is needed, subject-specific cross validation can be used.

There are different evaluation metrics for discrete emotions and continuous emotions. For discrete emotion recognition, it is a classification problem and the evaluation methods include accuracy of the recognition, F1-score (Koelstra et al., 2012), Receiver Operating Characteristic (ROC) curve and the area under the curve (Lucey et al., 2010), and confusion matrix (Kotsia and Pitas, 2007; Happy and Routray, 2015).

The accuracy of the recognition refers to the percentage of the correctly classified instances defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.1}$$

where TP = True Positive, FP = False Positive, TN = True Negative and FN = False Negative.

The F1 score is commonly used for the statistical analysis of binary classification. It measures the accuracy by considering both the precision ρ and the recall r, i.e., the harmonic mean of the precision and recall as shown in Equation 2.2. The best value of the F1 score is 1 and the worst is 0. The F1-score itself does not take the true negatives into account, therefore, the average of F1 score of the positive class and F1 score of negative class is usually used as the final evaluation metric.

$$F_1 = (\frac{recall^{-1} + precision^{-1}}{2})^{-1} = 2\frac{precision * recall}{precision + recall} \tag{2.2}$$

Confusion matrix is a specific table layout that allows visualisation of the performance of an algorithm, where each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). A typical problem encountered when evaluating the recognition results of affect recognition is the imbalanced data, which occurs when more samples are in one/several certain class(es) than the other class(es). The accuracy is not a reliable metric for evaluating the real performance of a classifier when the dataset is unbalanced, while confusion matrix is a good option in such cases.

Continuous emotion recognition is formulated as a regression problem and therefore is usually evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE) (Kollias et al., 2017), Pearson Correlation Coefficient (PCC) (Ringeval et al., 2015), and Concordance Correlation Coefficient (CCC) (Ringeval et al., 2015; Kollias et al., 2017; 2018).

MAE is a measure of differences between two continuous variables as shown in Equation 2.3, i.e., between estimated affective levels and the ground truth of the affective levels.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - \hat{X}_i| \tag{2.3}$$

where $\hat{X}$ is a vector of n predictions generated from a sample of n data points, and X is the ground truth values of the n data points.

MSE is used to measure the average of the squares of the errors as shown in Equation 2.4, which is the average squared difference between the estimated affective levels and the ground truth of the affective levels.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)^2 \tag{2.4}$$

The PCC (ρ) is a correlation measure between two variables, which is illustrated in Equation 2.5. It has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{2.5}$$

where cov is the covariance, σ_x is the standard deviation of x and σ_y is the standard deviation of y.

The CCC is also a correlation measurement between two variables, which combines the PCC with the square difference between the means of two compared variables as illustrated in Equation 2.6.

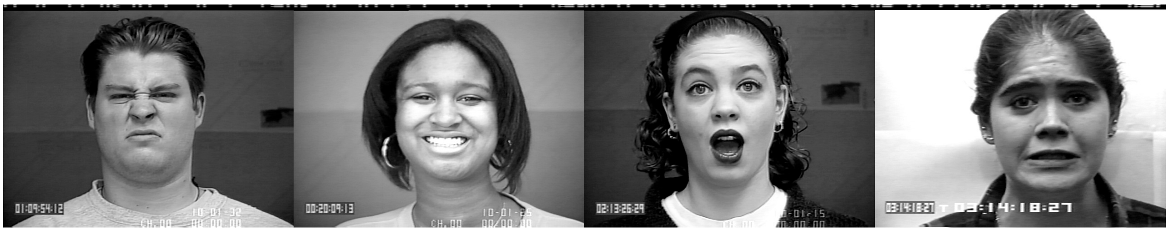
$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{2.6}$$

where ρ is the PCC between the ground truth and prediction, σ_x^2 and σ_y^2 are the variances, and μ_x and μ_y are the means of ground truth and prediction respectively.

2.4 Databases for affect analysis

2.4.1 Databases for affect analysis in individual settings

Most of the early databases for affect analysis in individual settings contain only posed expressions, which are usually collected in the controlled lab environment by asking the participants to pose different emotions. One of them is the Cohn-Kanade (CK) database (Kanade et al., 2000), published in 2001, and some example images of the posed emotions from this dataset are illustrated in Figure 2.3. Arguably, the CK database is one of the first and widely used datasets in the field of affect recognition (Kanade et al., 2000; Tian et al., 2001) and it contains 1,917 image sequences of only frontal faces and posed emotions with 182 subjects involved. The CK database was later extended to a database called CK+ (Lucey et al., 2010) by adding more posed emotions, some spontaneous emotions and some new subjects. One similar dataset to the CK database is the Japanese Female Facial Expression (JAFFE, published in 1999) database (Lyons et al., 1999), however, the database contains only 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. In addition to CK and CK+ databases, another important early database in affect analysis is the MMI Facial Expression Database (Pantic et al., 2005), which includes more than 1,500 samples of both static images and image sequences of faces displaying various facial expressions. It introduces some major improvements, such as adding profile views and temporal labelling of onset, apex and offset of emotions. The



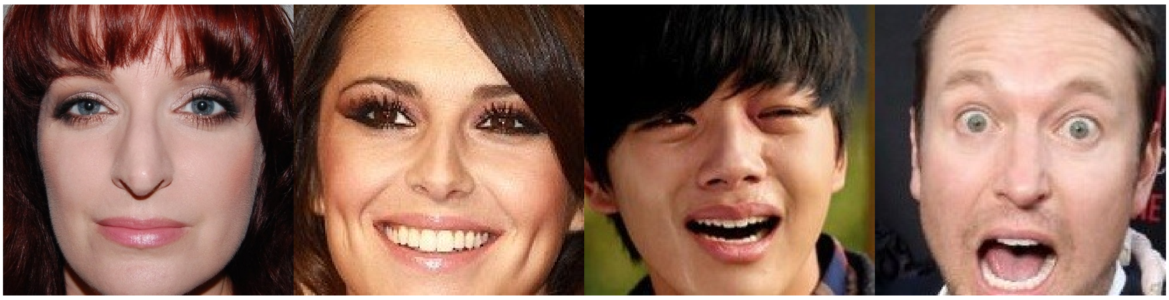
(a) Disgust. (b) Happy. (c) Surprise. (d) Fear.

Figure 2.3 Examples of the posed emotions from the CK database (Kanade et al., 2000). The corresponding emotions of the images are: (a) Disgust, (b) Happy, (c) Surprise, (d) Fear.

Multi-PIE (Gross et al., 2010) dataset further increases the data variability by including a very large number of views at different angles and various illumination conditions.

The aforementioned databases mainly contain posed emotions, however, the emotions humans display in the daily life can be very distinct from the posed emotions. Currently, it is widely accepted that the recognition of posed expressions, even though is an interesting research problem, is not very relevant for real world settings, where more subtle and complex emotions are usually displayed. Therefore, it is necessary to collect data for affect analysis in more naturalistic settings. To this end, a number of databases that include non-posed emotions and more spontaneous emotions have been collected in the past few years such as the FER-2013 (Goodfellow et al., 2013), Affectiva-MIT (McDuff et al., 2013) and AffectNet (Mollahosseini et al., 2017) databases. Some example images from the AffectNet database are shown in Figure 2.4. Compared to the posed emotions shown in Figure 2.3, the ones shown in Figure 2.4 can be seen as more naturalistic.

In addition, most of the early databases contain only unimodal signals, such as Multi-PIE (Gross et al., 2010), which has only visual images of faces. However, humans express emotions via various channels, e.g., facial expressions, speech and body gestures. Later on, some multi-modal datasets are available for the public, such as RECOLA (Ringeval et al., 2013) and DEAP (Koelstra et al., 2012). In the following sections, we will review various databases widely used in affect recognition from the posed to spontaneous ones and from



(a) Neutral (b) Happy (c) Sad (d) Surprise

Figure 2.4 Examples of the spontaneous emotions from the AffectNet database (Mollahosseini et al., 2017). The corresponding emotions of the images are: (a) Neutral, (b) Happy, (c) Sad, (d) Surprise.

unimodal to multi-modal ones. Table 2.1 reviews the widely used databases for automatic affect recognition.

AFEW and SFEW. The Acted Facial Expressions in the Wild (AFEW) and Static Facial Expressions in the Wild (SFEW) were collected from 54 movies by Dhall and colleagues (Dhall et al., 2012). The movie clips were selected based on the subtitles and then six basic emotions plus neutral were labeled by human annotators. AFEW contains 1,426 video sequences, while SFEW is a static subset of AFEW, i.e., SFEW consists of the selected frames from AFEW. SFEW aims to cover different head poses, various illuminations and occlusions while selecting frames from AFEW.

FER-2013. The Facial Expression Recognition 2013 (FER-2013) database was introduced in the ICML 2013 Challenges (Goodfellow et al., 2013). The database was created using the Google image search API that matched a set of 184 emotion-related keywords to capture the six basic expressions as well as the neutral expression.

EmotioNet. The EmotioNet dataset (Fabian Benitez-Quiroz et al., 2016) consists of one million images of facial expressions downloaded from the Internet by selecting all the words derived from the word "feeling" in the lexical database for English Word-Net (Miller,

Table 2.1 Databases used for affect analysis in individual settings

Database	Images/ videos	Posed/ Non-posed	Labels	Multi-modal or not
CK+ (Lucey et al., 2010)	Videos	Posed	-Basic emotions + contempt -30 AUs	No
Multi-PIE (Gross et al., 2010)	Images	Posed	Basic emotions + neutral	No
MMI (Pantic et al., 2005)	Images & videos	Posed & non-posed	Basic emotions + neutral	No
SFEW (Dhall et al., 2012)	Images	Non-posed	Basic emotions + neutral	No
AFEW (Dhall et al., 2012)	Videos	Non-posed	Basic emotions + neutral	No
RECOLA (Ringeval et al., 2013)	Videos	Non-posed	Valence-arousal (continuous)	-Yes -Video, audio, ECG and EDA
DEAP (Koelstra et al., 2012)	Videos	Non-posed	Valence-arousal (continuous)	-Yes -Videos and EEG signals
FER-2013 (Goodfellow et al., 2013)	Images	Non-posed	Basic emotions + neutral	No
Aff-Wild (images) (Zafeiriou et al., 2016)	Images	Non-posed	Basic emotions + neutral	No
Aff-Wild (videos) (Zafeiriou et al., 2016)	Videos	Non-posed	Basic emotions + neutral	No
Emotionet (Fabian Benitez-Quiroz et al., 2016)	Images	Non-posed	-12 AUs -23 emotion categories annotated based on AUs	No
AffectNet (Mollahosseini et al., 2017)	Images	Non-posed	-8 emotion categories -Valence-arousal (continuous)	No

Note: In the table, basic emotions refer to the six basic emotions defined by Ekman (Ekman and Friesen, 2003), i.e., “anger, disgust, fear, happiness, sadness, and surprise”.

1995). A total of 100,000 images were manually annotated with Action Units (AUs) [5] by experienced coders and the other images were automatically annotated with AUs and AU intensities. In total 23 categorical emotions were labelled based on AUs according to (Du et al., 2014), in which the emotion categories were defined based on AUs. EmotioNet is a large database with a large number of subjects of great variations. However, it lacks the dimensional model of affect, and the emotion categories are defined based on annotated AUs and not manually validated.

The aforementioned databases contain only categorical emotions. In addition to these datasets, a number of researchers collected databases with continuous dimensional annotations, such as Belfast (Sneddon et al., 2012), Aff-Wild (Zafeiriou et al., 2016) and AffectNet (Mollahosseini et al., 2017). Below we provide some details of the databases with continuous annotations.

DEAP. The Database for Emotion Analysis using Physiological Signals (DEAP) (Koelstra et al., 2012) collected spontaneous reactions of 32 participants in response to one-minute music videos. Both biological signals, i.e., electroencephalogram (EEG) and peripheral physiological signals, and frontal face videos of participants were recorded. Both valence and arousal annotations are provided in the DEAP dataset. Even though DEAP has a limited number of subjects and the videos were captured in lab controlled settings, it is a great database for affect analysis using different modalities.

Aff-Wild. Aff-Wild dataset (Zafeiriou et al., 2016) has two subsets for affect analysis, i.e., one subset of image data and one subset of video data. Aff-Wild dataset of videos is by far one of the largest databases for measuring continuous affect in the valence-arousal space "in-the-wild" in which more than 500 videos were collected. These videos were downloaded from YouTube and subjects in the videos displayed a number of spontaneous emotions while watching a particular video, performing an activity or reacting to a practical joke. The videos

[5] AUs are the atomic facial muscle actions defined by the Facial Action Coding System (FACS) (Friesen and Ekman, 1978). FACS is a system describing all visually discernible facial movement and breaks down facial expressions into individual components of muscle movement, i.e., AUs (Friesen and Ekman, 1978).

were annotated by three human annotators, utilising a joystick-based tool to rate along two continuous dimensions, i.e., valence and arousal.

AffectNet. By far, AffectNet (Mollahosseini et al., 2017) is the largest database for affect analysis in static images in the wild with both of the categorical and dimensional annotations. The database was created by querying emotion related keywords online and 450,000 images were manually annotated for the presence of eight emotion categories, i.e., six basic emotions, neutral and contempt, and the intensity of valence and arousal. As the aforementioned datasets in this subsection, there is only one person in each image in the AffectNet.

2.4.2 Databases for affect analysis in group settings

In this section, we will review currently available datasets for affect analysis in group settings. Compared to affect analysis in individual settings, affect analysis in group settings started later and there are a smaller number of databases available. We provide the details of each dataset in the following part of this section.

HAPPEI. The first database for group-level emotion analysis, named as HAPpy PEople Images (HAPPEI), was collected by Dhall *et al.* (Dhall, Joshi, Radwan and Goecke, 2012). This database contains 4,886 images that were collected from Flickr using key words, such as "party + people" and "graduation + ceremony". Each image was labelled with a group-level happiness intensity, face level happiness intensity, occlusion intensity and pose by four human annotators. The happiness intensity is categorised into six levels of happiness (0-5), i.e., neutral, small smile, large smile, small laugh, large laugh and thrilled. As the database name implies, the database only contains images of people with happy facial expressions, which are particularly useful for group happiness intensity analysis (Dhall, Goecke and Gedeon, 2015).

Figure 2.5 Examples of images from GAF dataset (Dhall, Joshi, Sikka, Goecke and Sebe, 2015). The upper, middle and lower rows show images with the a group of people display a positive, neutral and negative affect, respectively.

Group Affect Database. Dhall *et al.* also (Dhall, Joshi, Sikka, Goecke and Sebe, 2015) collected another database called Group Affect Database (GAF) containing 504 images, which extended the HAPPEI database from positive affect only (Dhall, Goecke and Gedeon, 2015) to other emotion categories, i.e., "positive, neutral and negative", but only along valence dimensions. Examples of images from the GAF database are shown in Figure 2.5. In the EmotiW challenge 2017 (Dhall et al., 2017), the GAF database was extended to GAF 2.0 that contains 6,471 images, which was labelled in the same way as Group Affect Database. In the EmotiW challenge 2018 (Dhall et al., 2018), GAF 2.0 was further extended to GAF 3.0 to contain 17,172 images and was labelled in the same way, i.e., "positive, neutral and negative".

MultiEmoVA. The aforementioned databases are either limited to happy emotions or limited to valence dimensions, i.e., "positive, neutral and negative". In contrast, Mou *et al.* (Mou et al., 2015) collected a dataset for group-level emotion analysis along both arousal and valence dimensions, which is named as MultiEmoVA. The MultiEmoVA database was collected using the key words, "meeting", "party", "conference", "group/people", "graduate ceremony", "crowd", "sports event" and "movies", from Flickr and Google Image. In total, 250 images with varying number of faces were selected after applying a face detector developed in (Xiong and De la Torre, 2013). Each image was annotated by 15 annotators and each annotator was asked to select one label from "low, medium, high" for arousal and one from "negative, neutral, positive" for valence, that best described the group-level emotion expressed by people in each image.

2.4.3 AMIGOS database

The AMIGOS database is used in this thesis, which was collected for multi-modal research of affect, personality traits and mood on individuals and groups (Correa et al., 2018). Different from the aforementioned databases, (1) it contains data in both individual and group settings and (2) it consists of dynamic videos, which have more information of group dynamics to explore than static images. In this thesis, the individual database (IndividualDB) and group database (GroupDB) belonging to AMIGOS dataset are used. In IndividualDB, the participant watches movies alone, while in GroupDB, participants watch movies in groups.

Data collection of AMIGOS. Before data collection, ethical approval was first obtained from the University, Queen Mary University of London. Then the advertisement of the experiment was sent to all of the residents in the university using the email list. Participants were recorded using a JVC GY-HM150E camera while watching emotional videos. Additionally, both RGB and depth body videos were recorded using a Microsoft's Kinect V15 placed at the top of the screen. In addition, the participants' physiological signals were recorded using wearable sensors. Electroencephalogram (EEG) signal was recorded using Emotiv EPOC Neuroheadset [6], Electrocardiogram (ECG) singal was recorded using the Shimmer 2R platform [7] and Galvanic Skin Response (GSR) was recorded using the Shimmer 2R platform. In this thesis, as we focus on vision-based affect analysis, only videos are used in the research. Further details about the database are provided in (Correa et al., 2018). In the individual setting, 17 participants performed the experiment alone. In the group setting, 16 participants performed the experiment in 4 groups with 4 people in each group. During the recording sessions, the participant(s) was(were) led to the recording room. Experimenters first explained the protocol and then the participants read and signed the information sheet and the consent forms. Experimenters avoided to mention whether the participants could talk

[6] http://www.emotiv.com/

[7] http://www.shimmersensing.com/

Table 2.2 The stimuli of long movies / videos are presented with their sources, movie types, durations, and IDs. In the remaining part of the thesis, the video IDs are used to refer to movies / videos.

Movie / Video	Movie type	Duration/ min	ID
The Descent. Dir. Neil Marshall. Lionsgate, 2005.	Horror	23:30	N1
Back to School Mr. Bean. Dir. John Birkin. Tiger Aspect Productions, 1994.	Comedy	18:38	P1
The Dark Knight. Dir. Christopher Nolan. Warner Bross, 2008.	Action	23:25	B1
Up. Dirs. Pete Docter and Bob Peterson. Walt Disney Pictures and Pixar Animation Studios, 2009.	Adventure	14:01	U1

during the experiment, for the interactions to be spontaneous. Once the sensors had been tested, the experimenters left the room and the recording session started.

During the recordings, the participants were asked to watch stimuli of different affective nature. In both databases, four long movie segments (14-24 mins) were used as movie stimuli, details of which are listed in Table 2.2 and snapshots from the four movies are shown in Figure 2.6. In IndividualDB, seventeen participants who were different from the sixteen participants in GroupDB, watched these four movies individually. Videos were recorded at 1280×720 resolution, 25fps.

Annotation of AMIGOS. The annotation was conducted by human labelers, three researchers whose research is focusing on affect analysis. Independent observer annotations were obtained by using one in-house affect annotation interface that requires the labelers to scroll a bar between a range of continuous values from -0.5 to 0.5. The labelers were asked to give one label for valence and one label for arousal for every 20 seconds starting from the beginning of each recording (e.g., the interval for 00:00∼00:20 min, and 00:21∼00:40 min.). The labeler annotated arousal and valence separately to avoid the confusion between these two dimensions; the 20-second recordings were played in a random order to each labeler; each labeler was asked to observe the visual behaviours without hearing any audio

(a) The snapshot from movie N1.

(b) The snapshot from movie P1.

(c) The snapshot from movie B1.

(d) The snapshot from movie U1.

Figure 2.6 Snapshots from the four movies used as stimuli in the AMIGOS dataset (Correa et al., 2018).

and rate a single annotation for each 20-second recording along either arousal or valence dimension. Each of the labelers annotated all of the video segments, which means that each video segment obtained three annotations from all of the three labelers.

The 2D graph of the annotated arousal-valence are shown in Figure 2.7 and 2.8 for both GroupDB and IndividualDB, where the blue ones represent GroupDB, while the red ones represent the IndividualDB. Figure 2.7 is for all of the clips from GroupDB and IndividualDB, while Figure 2.8 shows all groups for all videos (i.e., movies) separately. From Figure 2.7 and Figure 2.8, we can see that the distribution of emotions people express are different in IndividualDB and GroupDB. This is consistent with the findings of Hess, Banse and Kappas (Hess et al., 1995), i.e., watching multimedia content alone or together with others influences the intensity of people's emotions. In addition, we can also see that the variances of emotions expressed by people in group settings are larger than that of individual settings along both arousal (0.1152 for group settings and 0.0973 for individual settings in terms of standard deviation across all videos) and valence (0.0917 for group settings and 0.0781 for individual settings in terms of standard deviation across all videos) dimensions. The standard deviations of the annotated arousal and valence levels of different videos for individual and group settings are provided in Table 2.3. From Table 2.3, we can see that under all videos, the standard deviations of emotions expressed by people in group settings are larger than that of individual settings along both arousal and valence dimensions. From Figure 2.7, we can also see that both the high and low areas of the arousal and valence dimensions are covered. Figure 2.9 and 2.10 illustrate the emotions of different groups along time. From Figure 2.9 and 2.10, we can see that the emotions expressed by people in different groups have some differences.

In order to assess the inter-labeler agreement, Cronbach's α (Cronbach, 1951), that has been widely used in the literature for agreement assessment on continuous scale (Ringeval et al., 2013; Celiktutan and Gunes, 2014; Ringeval et al., 2015; Celiktutan and Gunes,

Table 2.3 Standard deviations of emotions in individual and group settings along arousal and valence for different videos.

Dimensions	Arousal				Valence			
	N1	P1	B1	U1	N1	P1	B1	U1
GroupDB	0.112	0.124	0.064	0.096	0.082	0.106	0.033	0.077
IndividualDB	0.077	0.120	0.042	0.066	0.046	0.098	0.026	0.048

Table 2.4 The results of the measurement of inter-labeler agreement in terms of Cronbach's α on the annotations are reported in terms of arousal and valence dimensions among 3 labelers.

Dimension	Arousal	Valence
Methods	Cronbach's α	Cronbach's α
GroupDB	0.80	0.89
IndividualDB	0.75	0.88

2017), were computed. Mean Cronbach's α over all participants for both GroupDB and IndividualDB are listed in Table 2.4. In the literature, the value of Cronbach's $\alpha > 0.7$ is considered as an acceptable agreement level and $\alpha > 0.8$ as a good agreement level (Ringeval et al., 2013; Celiktutan and Gunes, 2017).

To give an overview of the existing databases for affect analysis in group settings, the datasets mentioned above and AMIGOS are listed in Table 2.5 with their data sources, data types, number of images/frames, labels, number of external annotators and data collection settings.

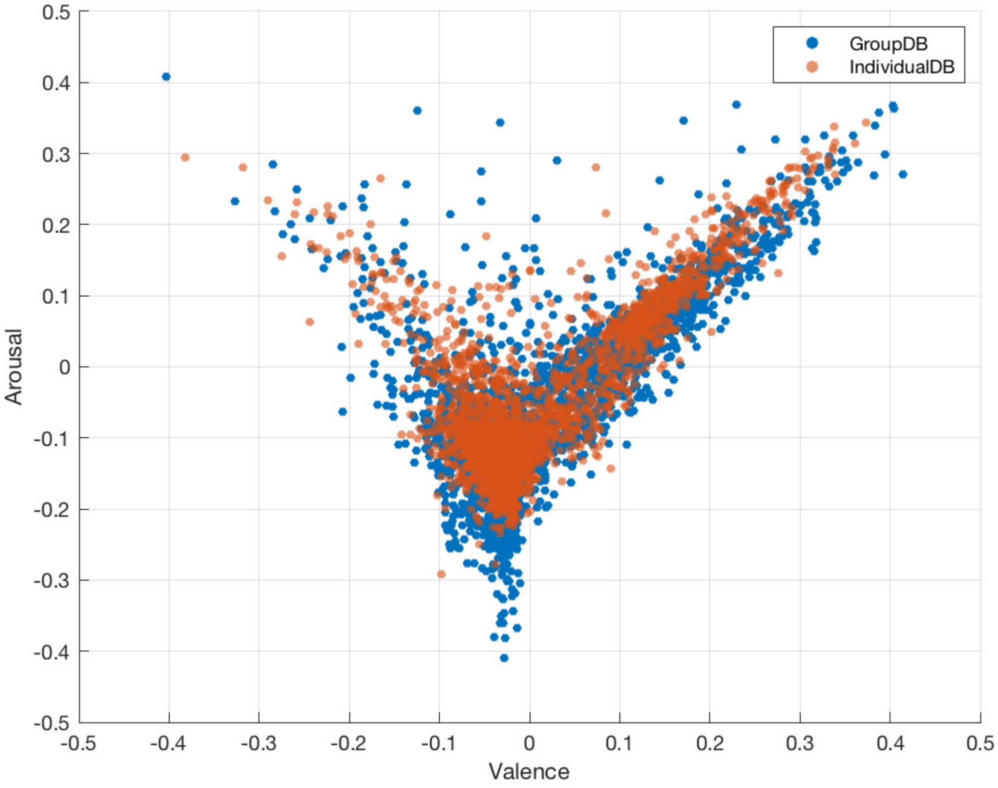
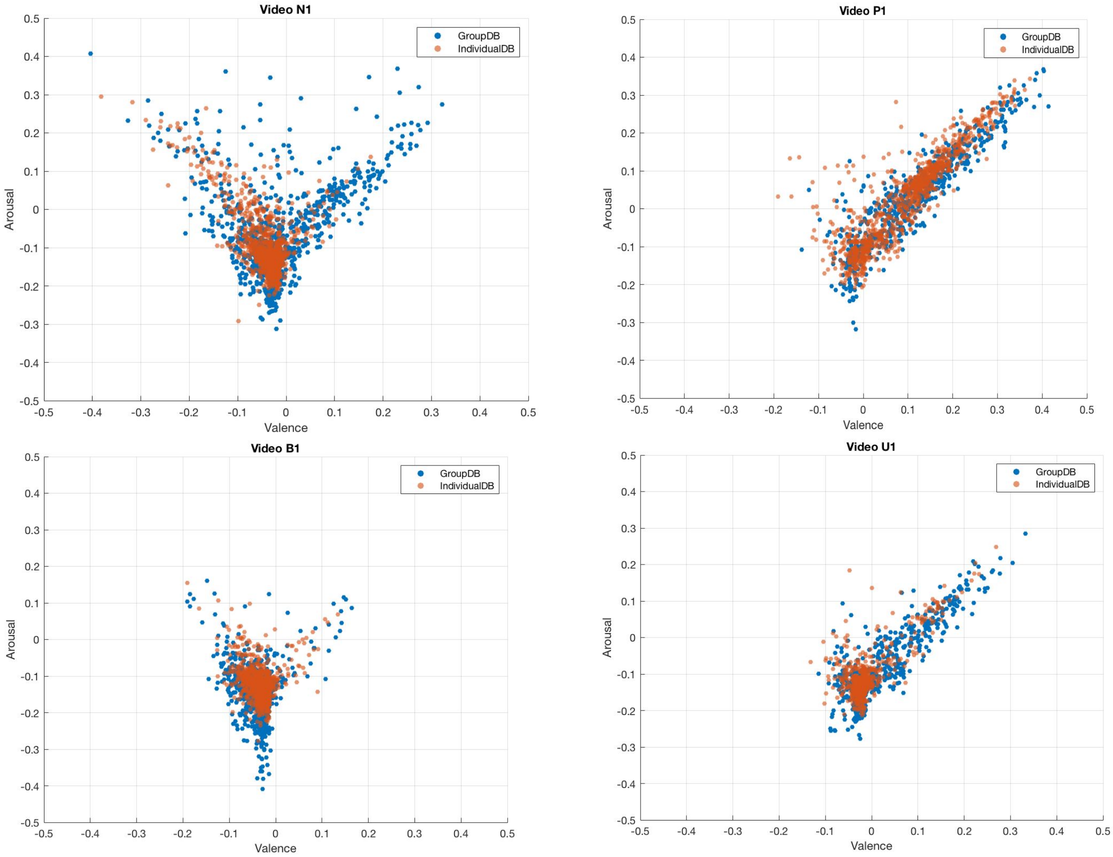


Figure 2.7 The affect annotation results for both GroupDB and IndividualDB along arousal and valence dimensions.



Video N1

Video P1

Video B1

Video U1

Figure 2.8 The affect annotation results for both GroupDB and IndividualDB along arousal and valence dimensions for all participants while watching different videos.

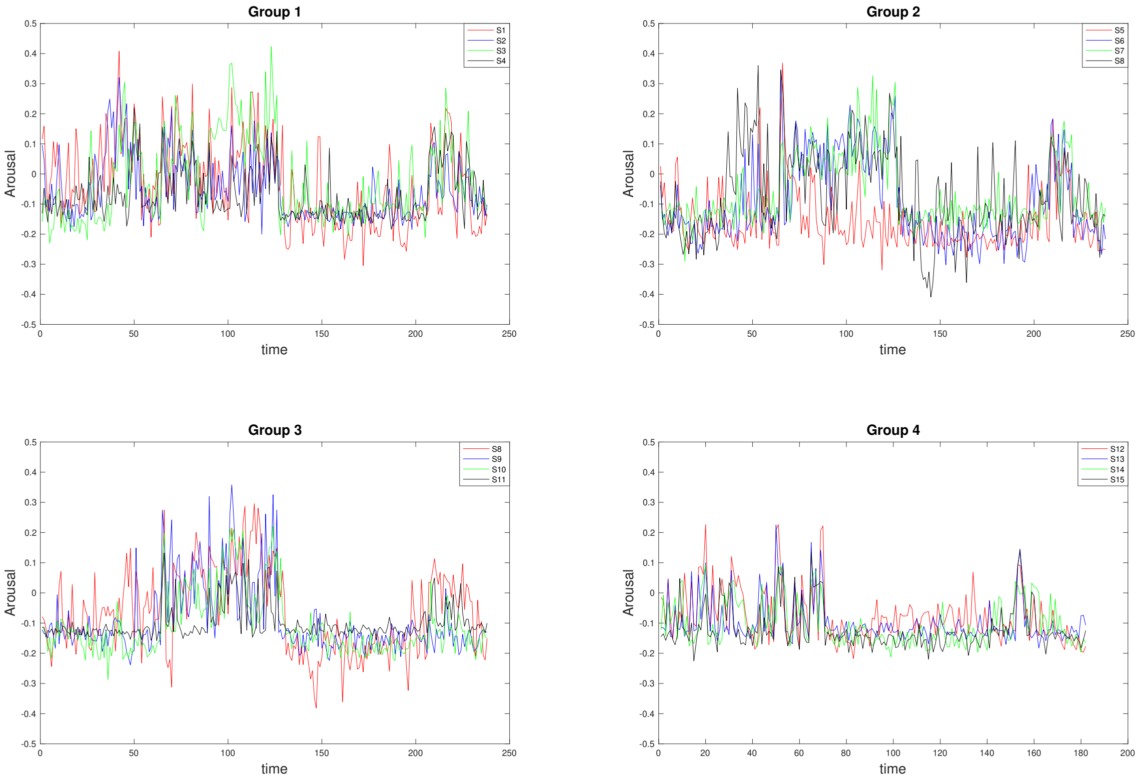


Figure 2.9 The annotated arousal levels for the four groups, one sub-figure for each group. The annotated arousal levels of four subjects in each group are plotted. The subject IDs, S1 to S16, are the same as those shown in Figure 2.11.

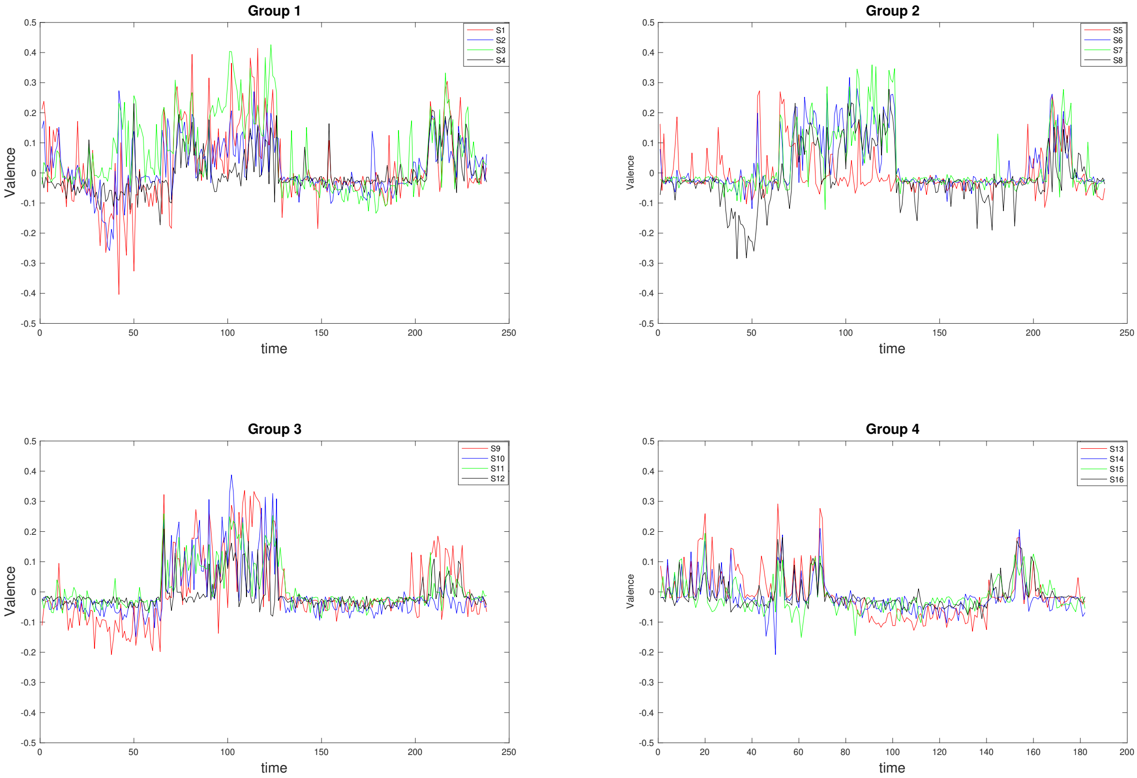
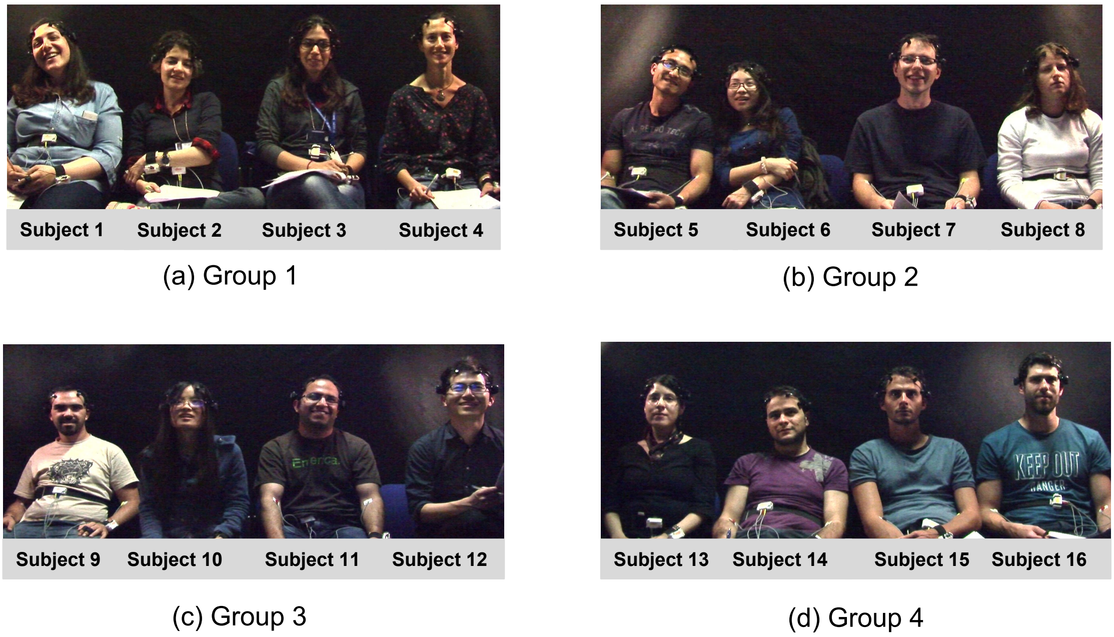


Figure 2.10 The annotated valence levels for the four groups, one sub-figure for each group. The annotated valence levels of four subjects in each group are plotted. The subject IDs, S1 to S16, are the same as those shown in Figure 2.11.

Table 2.5 An overview of databases on affect analysis in group settings

Database	HAPPEI (Dhall, Goecke and Gedeon, 2015)	GAF (Dhall, Joshi, Sikka, Goecke and Sebe, 2015)	GAF 2.0 (Dhall et al., 2017)	GAF 3.0 (Dhall et al., 2018)	MultiEmoVA (Mou et al., 2015)	AMIGOS (Correa et al., 2018)
Data Source	Web	Web	Web	Web	Web	Recordings
Data Type	Static	Static	Static	Static	Static	Dynamic
Num of images/ frames	4,886 (note that 3,134 images were used in the EmotiW 2016 challenges (Dhall et al., 2016))	504	6,471	17,172	250	More than 6,000,000 frames (12,580 short clips after segmentation)
Labels	6 stages of happiness (neutral, small smile, large smile, small laugh, large laugh and thrilled)	3 categories for valence (negative, neutral and positive)	3 categories for valence (negative, neutral and positive)	3 categories for valence (negative, neutral and positive)	3 categories for valence (negative, neutral and positive) and 3 categories for arousal (low, medium and high)	Self-assessment of valence, arousal, dominance and basic emotions. External annotation of continuous valence and arousal
Number of external annotators	4	3	——	——	15	3
Settings	Group	Group	Group	Group	Group	Individual & group



(a) Group 1

(b) Group 2

(c) Group 3

(d) Group 4

Figure 2.11 Four illustrative frames from four groups of data and the ID of each subject.

Chapter 3

RELATED WORK

Automatic affect recognition has received a lot of attention in recent years with various applications in very diverse areas such as human-computer interaction (Dautenhahn, 2007), security (Hernandez et al., 2012), healthcare (Kaltwang et al., 2012) and education (Kleinsmith and Bianchi-Berthouze, 2013). Most of the existing works on affect analysis have been carried out in *individual settings* where each individual stays alone (Poria et al., 2017; Gunes and Pantic, 2010). However, in the real world, people often stay with others, interacting in *group settings*. More recently an increasing number of works have started focusing on affect analysis in *group settings* and there are challenge events organised in this field since 2016 (Dhall et al., 2016; 2017; 2018). The literature review below introduces various feature representations in emotion recognition and works in affect analysis especially in *group settings*.

3.1 Features for affect analysis

Humans perceive and express emotions through many different channels such as visual, auditory and touching sensing. As this thesis focuses on affect analysis using visual signals, in this section, only the vision-based features are reviewed.

Vision-based information is the major one utilised for affect analysis (Zeng et al., 2009). Within these visual modalities, face is one of the most important channels of non-verbal communication and facial expressions have been one of the most prominent features in research for almost every aspect of emotion analysis. In the affective computing community, most of the research in vision-based emotion recognition has centred around facial expressions (Sariyanidi et al., 2015; Cohn and De la Torre, 2014; Karg et al., 2013; Kleinsmith and Bianchi-Berthouze, 2013). In addition, bodily expressions are also important for affect analysis (Kleinsmith and Bianchi-Berthouze, 2013; Gunes and Piccardi, 2007; Gunes et al., 2015; Huang et al., 2018). Even though we cannot find a unique relationship between a discrete emotion and a body expression, in the survey paper (Kleinsmith and Bianchi-Berthouze, 2013) it has been concluded that body information can be used for recognition of both continuous affective dimensions and discrete emotion categories. In this section, we will review the commonly used facial and body feature representations for affect analysis.

On one hand, the predesigned features for affect recognition can be divided into geometric and appearance features. Geometric features can represent the shape of the facial components, e.g., eyes and mouth, and the location of facial salient points, e.g., corners of the eyes and mouth (Pantic and Patras, 2006). For body, the geometric feature (i.e., body form) is also important for analysing affect (Kleinsmith and Bianchi-Berthouze, 2013). The postural configuration of arms and legs are geometric information of the body used for affect analysis (Kleinsmith and Bianchi-Berthouze, 2007). Appearance features represent the texture information (Sariyanidi et al., 2015). On the other hand, features can be split into static and dynamic features. The static features describe a single frame or image, while the dynamic features are able to encode the temporal information of videos or image sequences.

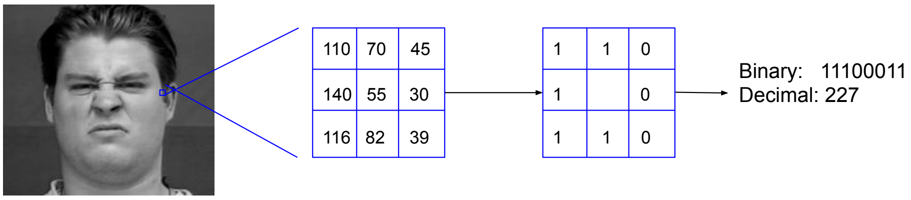


Figure 3.1 An illustration of how the standard LBP descriptor is extracted.

3.1.1 Appearance features for affect analysis

Some representative appearance features include Local Binary Patterns (LBP) (Ahonen et al., 2006), Local Phase Quantization (LPQ) (Ojansivu and Heikkilä, 2008), Histograms of Oriented Gradient (HOG) (Dalal and Triggs, 2005), and Quantized Local Zernike Moment (QLZM) (Sariyanidi et al., 2013). LBP was utilized by the winner of 2012 Audio-visual Emotion Challenge (AVEC) (Savran et al., 2012) and Facial Expression Recognition and Analysis (FERA) challenge (Yang and Bhanu, 2011); LPQ was used by prominent systems in FERA (Yang and Bhanu, 2011) and AVEC (Cruz et al., 2011); HOG based features are used to extract body features for affect recognition (Chen et al., 2013).

LBP operator (Ojala et al., 1996) is one of the best performing texture features and has been widely utilized in different applications. The standard LBP operator assigns a label to every pixel of an image by thresholding the 3×3-neighborhood of each pixel with the center pixel value and considering the result as a binary number. Then, the histogram of the labels can be used as a texture descriptor. Figure 3.1 illustrated how the standard LBP descriptor is extracted. The main advantages of LBP descriptor are its computational efficiency and its robustness to illumination changes, but it is not invariant to rotations. Since LBP descriptor is proposed, many variants of the LBP descriptor have been developed, such as uniform LBP (Ojala et al., 2002; Huang et al., 2011).

LPQ descriptor (Ojansivu and Heikkilä, 2008) utilizes local phase information by operating the Fourier phase locally over a M-by-M window at each image pixel. LPQ is a spatial blurring method and robust to image blurring. In digital image processing, it is well known that a convolution between the image and a point spread function (PSF) can be used to describe the spatial blurring. When it comes to the Fourier (i.e., frequency) domain, it results in:

$$G(u) = F(u) \cdot H(u) \tag{3.1}$$

where $G(u)$ is the Fourier transforms of the blurred image, $F(u)$ is the Fourier transforms of the original image and $H(u)$ is the Fourier transforms of the PSF. The phase information is a sum:

$$\angle G(u) = \angle F(u) + \angle |H(u)| \tag{3.2}$$

where $\angle G(u)$, $\angle F(u)$ and $\angle H(u)$ denotes the phase angle of $G(u)$, $F(u)$ and $H(u)$ respectively. If we assume PSF, $h(x)$ is centrally symmetric, i.e., $h(x) = h(-x)$, the Fourier transform $H(u)$ is real valued and as a result $\angle H(u)$ can be represented using a two-valued function:

$$\angle H(u) = \begin{cases} 0 & \text{if } H(u) \geqslant 0 \\ \pi & \text{if } H(u) < 0 \end{cases}$$

This means that for all $H(u) >= 0$, $\angle G(u) = \angle F(u)$.

A short-term Fourier transform (STFT) is computed over a local M × M neighbourhood N_x at each pixel position x of the image $f(x)$:

$$F(u,x) = \sum_{y \in N_x} f(x-y) e^{-j2\pi u^T y} \tag{3.3}$$

In LPQ, only 4 complex coefficients are used, $u_1 = [a,0]^T$, $u_2 = [0,a]^T$, $u_3 = [a,a]^T$, $u_4 = [a,-a]^T$, where a is a scalar to satisfy $H(u_i) >= 0$. For each pixel x, we have:

$$F(x) = [Re\{F(u_1,x)\}, Im\{F(u_1,x)\}, ...Re\{F(u_4,x)\}, Im\{F(u_4,x)\}] \tag{3.4}$$

The phase information in the Fourier coefficients is recorded by observing the signs of the real and imaginary parts of each component in $F(x)$ using a simple scalar quantizer:

$$q_j = \begin{cases} 1 & \text{if } f_j \geqslant 0 \\ 0 & \text{if otherwise} \end{cases}$$

where f_j is the j_{th} component of $F(x)$. In this way, the eight binary coefficients q_j can be converted to an integer value between 0-255 using binary coding:

$$f_{LPQ} = \sum_{j=1}^{8} q_j 2^{j-1} \tag{3.5}$$

HOG descriptor represents images using the directions of the edges that the images contain. To extract HOG features, first, the image of face or body is divided into a number of blocks containing cells, where gradient magnitude and angle are extracted for each pixel. The local histogram of each cell is calculated using the gradient magnitude and angle/direction. A bin is selected based on the gradient angle/direction, and the vote (i.e., the value that goes into the bin) is selected based on the gradient magnitude. Then the local histograms from the cells are combined across the blocks to form the histogram representation of the image. It is possible that the blocks have overlaps with each others. The details are illustrated in Figure 3.2. In this manner, the dimensionality of the HOG descriptor for the image equals to $N_{blocks} \times N_{cells} \times N_{bins}$, where N_{blocks} refers to the number of blocks of each image; N_{cells} denotes the number of cells of each block; and N_{bins} denotes the number of bins of the histogram of one cell.

In addition, Quantised Local Zernike Moments (QLZM) is another low-level appearance feature in a histogram representation. It describes a neighbourhood by using its Local Zernike Moments (Sariyanidi et al., 2013). Each moment coefficient provides information of the variation at a unique scale and orientation, and there are no overlaps for the information

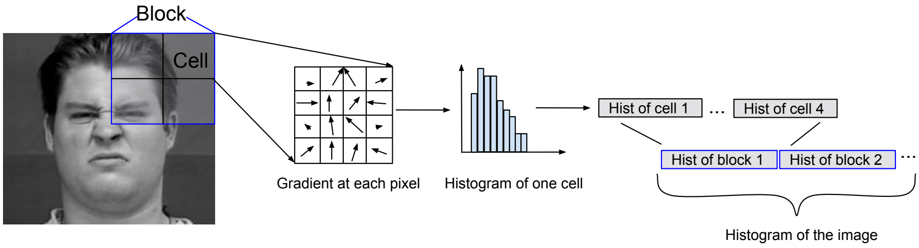


Figure 3.2 An illustration of how HOG descriptor is extracted.

conveyed by different moment coefficients (Sariyanidi et al., 2015). The QLZM descriptor is obtained by quantising all moment coefficients into an integer, and then converting to histograms.

Zernike Moments. The Zernike moment(ZM) consists of a complete orthogonal system in a unit circle, and is characterised by having size invariant to rotation. In addition, compared to the other moment quantities, it is also reported as being robust with respect to noise. Let $I(x,y)$ be the input image of size $X \times Y$. Zernike Moments (ZMs) are computed by decomposing $I(x,y)$ onto ZM basis matrices, a set of complex matrices that are orthogonal on the unit disk. Let the basis matrices be denoted with V_{nm} and defined through the radial polynomials R_{nm} as (Sariyanidi et al., 2013):

$$V_{nm}(\rho,\theta) = R_{nm}(\rho)e^{im\theta} \tag{3.6}$$

where ρ and θ are the radial coordinates, n is the order of the polynomial that controls the number of coefficients and m is the number of iterations, which can be set to any value so that $|m| < n$ and $n - |m|$ is even. R_{nm} are the radial polynomials, defined as (Sariyanidi et al., 2013):

$$R_{nm} = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s \rho^{n-2s}(n-s)!}{s!(\frac{n+|m|}{2}-s)!(\frac{n-|m|}{2}-s)!} \tag{3.7}$$

where x and y are the image coordinates mapped to the range $[-1,+1]$, $\rho_{xy} = \sqrt{\bar{x}^2+\bar{y}^2}$, and $\theta_{xy} = tan^{-1}\frac{\bar{y}}{\bar{x}}$. A ZM coefficient of $I(x,y)$, Z^{I}_{nm}, consists of a real and an imaginary component and can be computed as follows:

$$Z^{I}_{nm} = \frac{n+1}{\pi}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1} I(x,y)V^{*}_{nm}(\rho_{xy},\theta_{xy})\Delta\bar{x}\Delta\bar{y} \tag{3.8}$$

Note that the basis matrices V_{nm} are generic and do not depend on the input image. Local ZMs are computed from N local blocks, I_N, rather than the entire image $I(x,y)$. ZM coefficients are scattered in a wide range when computed globally but can be concentrated in a short range when computed locally.

Since a local descriptor represents the discontinuities and texture of an image effectively, QLZM is proposed in (Sariyanidi et al., 2013) using non-linear encoding, which facilitates the relevance of low-level features by increasing their robustness against image noise.

Non-linear encoding is carried out on complex-valued local ZMs using binary quantisation, which converts the real and imaginary parts of each ZM coefficient into binary values using $signum()$ functions. Specifically, let $Z^{I_N} = [Z^{I_N}_{p_1q_1},\ldots,Z^{I_K}_{p_Kq_1}]$ be a vector of K complex ZMs of I_N, and the complex notation of each coefficient be $Z^{I_N}_{pq} = Z^{I_N,\Re}_{pq} + iZ^{I_N,\Im}_{pq}$. We compute Q^{I_N}, the vector of quantised local ZM coefficients as follows:

$$Q^{I_N} = [Q^{I_N,\Re}_{p_1q_1}, Q^{I_N,\Im}_{p_1q_1},\ldots,Q^{I_N,\Re}_{p_Kq_K}, Q^{I_N,\Im}_{p_Kq_K}]_{1\times 2K'} \tag{3.9}$$

where $Q^{I_N,\Re}_{p_iq_i} = signum(Z^{I_N,\Re}_{p_iq_i})$. However, the basis matrices V_{nm} must be zero-mean to ensure that the output of $sgn()$ applied to coefficients computed through equation 3.8 is not biased. For any $m \neq 0$, this can be easily shown by computing the integral of V_{nm} over ρ and θ. For the continuous case ($\theta \in \Theta, \rho \in P; \Theta = [\pi,\pi], P = [0,1]$), it can be shown that:

$$\begin{aligned} \iint\limits_{\Theta,P} V_{nm}(\rho,\theta)d\rho d\theta &= \int\limits_{\Theta} e^{im\theta} \overbrace{\int\limits_{P} R_{nm}(\rho)d\rho}^{C(P)} d\theta = C(P)\int\limits_{-\pi}^{\pi} e^{im\theta}d\theta \\ &= \frac{C(P)}{im}[e^{im\theta}]_{-\pi}^{\pi} = \frac{C(P)}{im}[(cos\theta+isin\theta)^{m}]_{-\pi}^{\pi} \end{aligned} \tag{3.10}$$

On the other hand, for m = 0 it can be shown that $\iint\limits_{\Theta,P} V_{nm}(\rho,\theta)d\rho d\theta = 2\pi C(P)$, i.e., the mean of basis matrices is not zero for $C(P) \neq 0$. Therefore, we neglect the ZM coefficients with $m = 0$ while extracting local ZMs. Following the general rule of ZMs ($|m| < n$ and $n|m|$), we select local ZM coefficients such as $Z^{I_N} = [Z_{11}^{I_N}, Z_{22}^{I_N}, Z_{31}^{I_N}, Z_{33}^{I_N}, ...]_{1\times K}$ and the QLZM vector becomes $Q^{I_N} = [Q_{11}^{I_N,\Re}, Q_{11}^{I_N,\Im}, Q_{22}^{I_N,\Re}, Z_{22}^{I_N,\Im}, ...]_{1\times 2K}$. The number of moment coefficients, K, can be considered as a function of n and is computed as shown in equation 3.11. The size of each local histogram is 2^{2K}, and the length of the final vector for each image will depend on how many local blocks the image is divided into, where K is from equation 3.11.

$$K(n) = \begin{cases} \frac{(n+1)^2}{4} & \text{if } n \text{ is odd} \\ \frac{n(n+2)}{4} & \text{if } n \text{ is even} \end{cases} \tag{3.11}$$

In addition, a partitioning is also applied as shown in Figure 3.3. The final QLZM feature is constructed by concatenating all local histograms, and the length of extracted correspond to two parameters, i.e., the number of moment coefficient K (K is computed using the function of moment order n as shown in equation 3.11.) and the size of the grid M, computed as:

$$2^{2K} \times [M^2 + (M+1)^2] \tag{3.12}$$

Compared to LBP and LPQ, QLZM can be tuned to obtain lower-dimensional histograms. For example, LBP is in a dimension of 2478 (Shan and Gritti, 2008; Shan et al., 2009), and LPQ is in a dimension of 2048 (Jiang et al., 2011), while QLZM in a dimension of 656 showed very good performance in terms of facial expression recognition (Sariyanidi

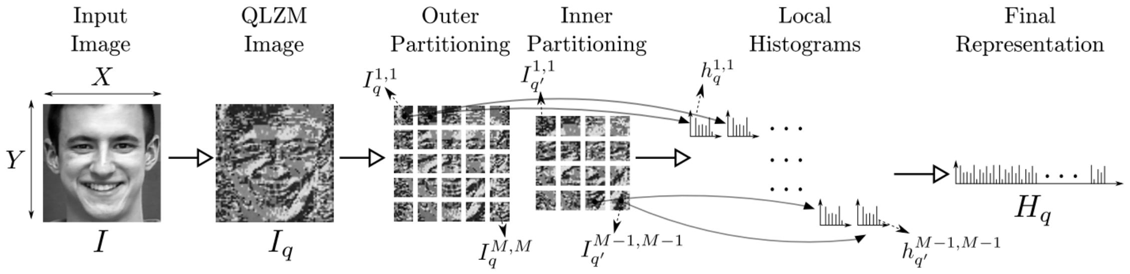


Figure 3.3 Illustration of the extraction of QLZM based facial representation framework (Sariyanidi et al., 2013).

et al., 2013). As QLZM achieved the state-of-the-art for affect recognition at the time of publication (Sariyanidi et al., 2013) and has a relatively lower dimension, in this thesis, we extend the static QLZM to volume representation to encode spatio-temporal information.

3.1.2 Geometric features for affect analysis

In addition to appearance features, geometric features that preserve the geometric information of the face or body are also used for affect recognition.

Geometric features can describe faces and bodies through distances and shapes, which can be distances between facial landmarks points (Kaya et al., 2015) or deformation parameters of a mash model (Kotsia and Pitas, 2006) and distances between body joints (Piana et al., 2013). How the geometric properties of the faces or bodies changing over time is also one of the important dynamic cues. To encode the dynamic information of the face and body, motion information are often estimated from color or intensity information, which can be extracted through optical flow (Wöllmer et al., 2013), which is described in Subsection 3.1.3. In (Afshar and Ali Salah, 2016; Kaya et al., 2015), after the facial landmark points are detected, geometric features are extracted by calculating the distances and angles between certain landmark points. For example, the distance between mouth points is used to show the mouth opening and the angle between the points of eyebrow is used to show the eyebrow slope.

In (Jung et al., 2015), to use the geometry information for affect recognition, the facial landmarks are first detected and then the normalized facial landmarks are taken as descriptors and fed into a deep neural network. The facial landmarks for one face at frame t are shown here:

$$X^{(t)} = [x_1^{(t)}, y_1^{(t)}, x_2^{(t)}, y_2^{(t)}, ..., x_n^{(t)}, y_n^{(t)}] \tag{3.13}$$

where n denotes the number of facial landmarks at frame t, $(x_i^{(t)}, y_i^{(t)})$ denotes the coordinate of the i-th facial landmarks at frame t. Then $(x_i^{(t)}, y_i^{(t)})$ is normalized by first subtracting the nose position and then divided by the standard deviations of all facial landmarks in each frame as follows:

$$\bar{x}_i^t = \frac{(x_i^t - x_{nose}^t)}{\sigma_x^t}, \bar{y}_i^t = \frac{(y_i^t - y_{nose}^t)}{\sigma_y^t} \tag{3.14}$$

Therefore, the final descriptor that is fed into the network is:

$$\bar{X}^{(t)} = [\bar{x}_1^{(t)}, \bar{y}_1^{(t)}, \bar{x}_2^{(t)}, \bar{y}_2^{(t)}, ..., \bar{x}_n^{(t)}, \bar{y}_n^{(t)}] \tag{3.15}$$

Geometric features are effective for affect recognition and not sensitive to the illumination changes, however, compared to appearance features, geometric features are not good at capturing the subtle textures, such as wrinkles and frowns on the face. In this thesis, we explore both geometric features and appearance features.

3.1.3 Temporal features for affect analysis

By now we have discussed various feature representations for affect recognition, both facial features and body features. However, the aforementioned representations are all features extracted from static images or standalone frames and do not contain any temporal information. Most of these methods are not working well with dynamic videos due to the lack of temporal information (Poria et al., 2017). In order to address this issue, researchers have proposed various methods to encode temporal information into the extracted features.

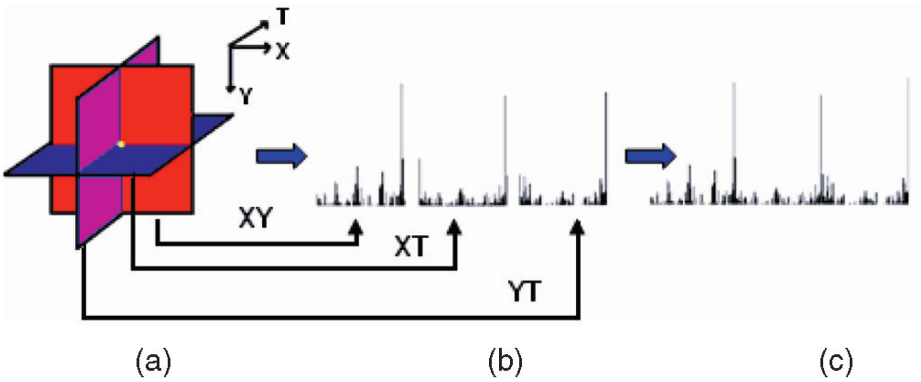


Figure 3.4 An illustration of one spatio-temporal feature, LBP-TOP (Zhao and Pietikainen, 2007). (a) Three orthogonal planes, i.e., XY, XT, and YT. (b) LBP histogram extracted from each plane. (c) Concatenate feature histograms from the three orthogonal planes.

Three Orthogonal Planes (TOP) methods. One of the important approaches is to extract low-level features from Three Orthogonal Planes (TOP) and then concatenate features of these three planes to form one representation for a video sequence. This approach was first proposed by Zhao *et al.* (Zhao and Pietikainen, 2007), which extended the static LBP feature to LBP-TOP feature for extracting spatial-temporal information. Figure 3.4 illustrates how LBP-TOP is extracted from three orthogonal planes, i.e., XY, XT and YT. LBP-TOP is used for discrete emotion recognition and shows good performance (Zhao and Pietikainen, 2007; Zhao and Pietikäinen, 2009). Following the success of LBP-TOP, the LPQ feature is also extended in the same way to the LPQ-TOP feature that has been used for AU detection and temporal segment recognition (Jiang et al., 2011; 2014; Afshar and Ali Salah, 2016).

Volume-based methods. The other way to encode spatio-temporal information is to extend the static features to volume representations, such as HOG 3D (Klaser et al., 2008). Given an interest region r_s as shown in Figure 3.5 (a), a descriptor d_s is used to represent the region by a feature vector. The interest region r_s is first divided into a set of $M \times M \times N$ cells, where one cell is denoted as c_i as shown in Figure 3.5 (a). For each c_i, it is divided into $S \times S \times S = S^3$ sub-blocks, b_j. The histogram for each cell is obtained by summing the histograms of all sub-blocks:

$$h_i = \sum_{j=1}^{S^3} q_j \tag{3.16}$$

q_j is a n-bin histogram of gradient orientations for a sub-block, b_j. If it is in 2D space, a n-bin histogram of gradient orientation can be seen as approximation of a circle with a regular n-sided polygon, where each side of the polygon corresponds to a histogram bin. In a similar manner, when it comes to the 3D space, the polygon in 2D space becomes a polyhedron as shown in Figure 3.5 (c). There are only five regular polyhedrons with congruent faces (these kind of polyhedrons are called platonic solids), i.e., the tetrahedron (4-sided), cube (6- sided), octahedron (8-sided), dodecahedron (12-sided), and icosahedron (20-sided). The authors of HOG 3D considered dodecahedron, and icosahedron for 3D gradient quantization as shown in Figure 3.5. The dimensionality of q_j, i.e., the number of orientations (nOrientation), depends on the option of the quantization. For example, if dodecahedron is used, nOrientation$=12$, while if icosahedron is used, nOrientation$=20$. In this manner, the dimensionality of the final descriptor, d_s is $M \times M \times N \times nOrientation$. HOG 3D is originally proposed for action recognition (Klaser et al., 2008), but applied to emotion recognition later on and the recent published papers on affect recognition still compare the results with that obtained using HOG 3D (Jung et al., 2015; Yang, Ciftci and Yin, 2018). HOG 3D is also used in our experiments for affect recognition and compared with other features.

Motion-based methods. As emotions are displayed over time, motion based features have been used to obtain temporal information from both the face and body. The optical flow method (Horn and Schunck, 1981) is one of the most important techniques that have been successfully applied to emotion recognition to extract dynamic movement in image sequences (Sidavong et al., 2019; Gunes and Piccardi, 2005). Optical flow is defined as an apparent motion of image brightness. If $I(x, y, t)$ is the intensity of a pixel at location (x, y) on frame I_1 at time t and it is offset by the flow (u, v) at time $t+1$ on frame I_2. Then the

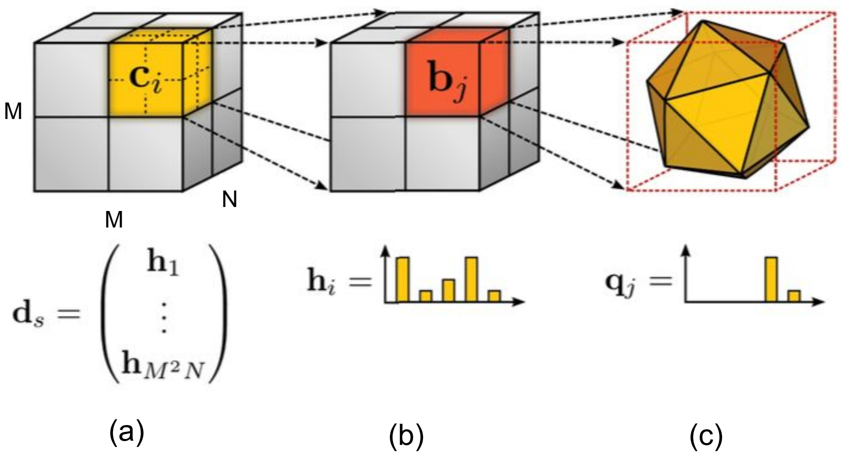

M
M
N
$\mathbf{c}_i$
$\mathbf{b}_j$
$\mathbf{d}_s = \begin{pmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{M^2N} \end{pmatrix}$
$\mathbf{h}_i =$
$\mathbf{q}_j =$
(a)
(b)
(c)

Figure 3.5 An illustration of how the HOG 3D descriptors are extracted (Klaser et al., 2008).

brightness constancy is given by:

$$I(x,y,t) = I(x+u, y+v, t+1) \tag{3.17}$$

$$I(x,y,t) = I(x,y,t) + u\frac{\delta I}{\delta x} + v\frac{\delta I}{\delta y} + \frac{\delta I}{\delta t} \tag{3.18}$$

$$0 = u\frac{\delta I}{\delta x} + v\frac{\delta I}{\delta y} + \frac{\delta I}{\delta t} \tag{3.19}$$

$$I_x = \frac{\delta I}{\delta x}, I_y = \frac{\delta I}{\delta y}, I_t = \frac{\delta I}{\delta t} \tag{3.20}$$

$$I_x u + I_y v + I_t = 0 \tag{3.21}$$

The above equation can be rewritten as:

$$\Delta I \cdot \overrightarrow{V} = -I_t \tag{3.22}$$

Where ΔI is the spatial intensity gradient and $\overrightarrow{V}$ is optical flow of pixel (x,y) at time t. The brightness constancy equation only defines the gradient of the moving pixels, but the boundaries of motion remain obscure, therefore, to estimate the actual flow, additional constrains are needed, e.g, Lucas-Kanade method (Lucas et al., 1981).

Mase was one of the first to start using optical flow for recognising facial expressions, where optical flow is used to determine the main direction of the movement of facial muscles. In (Otsuka and Ohya, 1997), optical flow is first obtained and then the 2D Fourier Transform coefficients of the optical flow are computed and used as the descriptors for facial expression recognition. It is not only for faces, optical flow is also used to extract body motion information for affect analysis (Gunes and Piccardi, 2008). Recently, dense trajectories are proposed for action recognition and achieved the-state-of-the-art recognition results before the use of deep learning based methods (Wang et al., 2013), where the trajectories are tracked using an optical flow method. As dense trajectories based methods achieved the-state-of-the-art performance for action recognition (Wang et al., 2013) due to the good capability of encoding spatio-temporal information, it is explored to be used for affect recognition in this thesis.

3.2 Affect analysis in group settings

Until now, we have discussed the different aspects of affect analysis, including datasets, features, and evaluation metrics, that can be applied to both individual settings and group settings. We also discussed some perspectives on the differences of these in both settings. In this subsection, we will focus on further discussions and insights into affect analysis in group settings, highlighting new challenges brought by the dynamic interactions between group members, approaches carefully designed for the scenarios with multiple persons involved and other related topics in group settings.

3.2.1 An overview of affect analysis in group settings

In the early years of affect analysis, most of the works focused on individual settings. However, social psychologists have shown that groups hold attributes that go beyond and exist in addition to the individual attributes of group members, and therefore a group is more than the sum of its parts (Bon, 1896; Sandelands and Clair, 1993). Preliminary works also showed that the social context (watching videos alone or in a group) affects the Quality of the Viewing Experience (QoVE) in terms of five aspects, i.e., enjoyment, endurability, satisfaction, involvement in the viewing experience and perceived visual quality (Zhu et al., 2014). In terms of emotions displayed in group settings, on one hand, individuals contribute their own individual feelings to the group and thus shape the emotion of the whole group; and on the other hand, the emotion of the group affect the individuals within a group and infuse them with distinct feelings (Menges and Kilduff, 2015; Barsade and Gibson, 2012). People reported both quantitatively and qualitatively different emotions in terms of joy, fear, anger and so forth while they were in different settings, i.e., in an individual setting or in a group setting (Mackie and Smith, 2017). Miranda-Correa *et al.* (Correa et al., 2018) have shown that the affect expressed by individuals heavily depends on the social context, i.e., whether an individual is alone or in a group. Consequently, emotions of each individual displayed in a group setting may differ from that of an individual expressed in an individual setting due to the influences of various factors in group settings. Therefore, we can investigate emotions in group settings from two perspectives: (1) emotion analysis of the whole group, i.e., group-level emotion analysis and (2) emotion analysis of each individual in group settings, i.e., individual-level emotion analysis.

Group-level affect refers to the affect displayed by the whole group of people collectively in an image or in a video, which has attracted a number of researchers to investigate in recent years. Group-level affect analysis in group settings can be conceptualised into two different ways, using a bottom-up approach and a top-down approach (Barsäde and Gibson, 1998). The bottom-up approach uses the affect of each individual in the group to obtain the group-level affect of the whole group (Dhall, Goecke and Gedeon, 2015). Hernandez *et al.* (Hernandez et al., 2012) conducted an interesting experiment, wherein the facial expression of the people passing through the corridor was analysed for the presence of smile. The number of smiles was averaged at a given point to decide the group-level mood. The top-down approach states that group-level emotions may arise from social identity or group membership, and are socially shared within a group and influence the emotions of each individual group member. Dhall *et al.* proposed a framework to infer the *overall happiness mood intensities* conveyed by a group of people in static images in (Dhall, Goecke and Gedeon, 2015) by combining the top-down and the bottom-up approaches. The winner papers (Li et al., 2016; Tan et al., 2017; Guo et al., 2018) of group affect sub-challenge in EmotiW challenges series (Dhall et al., 2016; 2018; 2017) also used both top-down and bottom-up approaches. In addition, Dhall *et.al* introduced a framework to predict the collective valence level of a group (i.e., positive, neutral and negative) in (Dhall, Joshi, Sikka, Goecke and Sebe, 2015) using both top-down and bottom-up approaches.

Individual-level affect analysis in group settings has been paid less attention to, compared to group-level affect analysis. To the best of our knowledge, the only work centers on individual-level affect recognition along valence and arousal dimensions in group settings is by Miranda-Correa *et al.* (Correa et al., 2018). They explored the individual-level affect recognition in group settings using physiological signals, i.e., EEG, ECG and GSR, along valence and arousal dimensions. A Gaussian Naïve Bayes classifier was used and in addition to the single modality, the fusion of the three types of signals were also reported. However, they only investigated how to recognise individual affect in group settings using physiological signals without considering any visual information. To fill this research gap, in this thesis we propose methodologies for individual-level affect analysis in group settings along valence and arousal dimensions based on visual information, as detailed in Chapter 4 and Chapter 5.

Table 3.1 provides an overview of the works on affect analysis in group settings, both in group-level and in individual-level, in terms of the utilised features, modalities, and methodologies etc.

3.2.2 Multi-modal affect analysis in group settings

Humans understand and express emotions through various channels, including faces, body gestures, audio etc., therefore, in terms of automatic affect recognition, it is important to investigate multi-modal frameworks. It has been demonstrated that the multi-modal framework can outperform a uni-modal approach for affect analysis in both individual and group settings (Poria et al., 2017; Dhall, Joshi, Sikka, Goecke and Sebe, 2015; Mou et al., 2015; Kahou et al., 2016). Compared to individual settings, multi-modal framework is more important in group settings due to the complex situations that involve multiple people and various dynamics. For example, as there are multiple people in an image or in a video, it is very common to have occlusions due to moving people. In this case, single modality, e.g., face or body, may not be always visible, thereby, it is important to utilise both face and body information. Studies have shown that the displayed affect heavily depends on context, such as where the person is and what the person is doing at that time (Vlachostergiou et al., 2014). Therefore, in addition to the face and body information, using context information is becoming increasingly popular for automatic affect recognition (Morency, 2013). The context information can be of great importance especially for group settings with multiple people that inherently involve complex contextual situations. In group settings, the context refers to not only each individual's identity, location and task but also their interpersonal dynamics, e.g., who the person is with and what others are doing at that time. The contextual information based on the group structure was used to infer group-level affect in (Mou et al., 2015) and individual gender and age in (Gallagher and Chen, 2009); and scene contextual

Table 3.1 Representative works on affect analysis in group settings

Refs	Data	Task	Channels	Hand-crafted / Deep-based features	Features
(Dhall, Goecke and Gedeon, 2015)	HAPPEI	Group-level happiness intensity recognition	-Face	Hand-crafted	PHOG
(Huang et al., 2015)	HAPPEI	Same as above	-Face	Hand-crafted	RVLBP
(Li et al., 2016)	HAPPEI	Same as above	-Face -Scene	Deep-based & Hand-crafted	-Face: ResNet-18 -Scene: CENTRIST
(Mou et al., 2015)	MultiEmoVA	Group-level affect recognition	-Face -Body -Context	Hand-crafted	-Face: QLZM -Body: HOG -Context: relative location and scale of each face to the image
(Dhall, Joshi, Sikka, Goecke and Sebe, 2015)	GAF	Group-level affect recognition of positive, neutral and negative	- Face - Scene	Hand-crafted	-Face: low-level LPQ and PHOG; higher-level action units -Scene: CENTRIST descriptor
(Tan et al., 2017)	GAF	Same as above	-Face -Scene	Deep-based	-Face: CNN based -Scene: CNN based
(Guo et al., 2018)	GAF	Same as above	-Face -Scene -Skeleton	Deep-based	-Face: VGG model -Scene: Inception-V2 & SE-ResNet-50 -Skeleton: SE-ResNet-50
(Huang et al., 2018)	HAPPEI & GAF	Group-level happiness intensity and affect	-Face -Upper body -Scene	Hand-crafted	-Face: RVLBP -Body: PHOG & LBP
(Ghosh et al., 2018)	HAPPEI & GAF	Group-level happiness intensity and affect	-Face -Scene	Deep-based	-Face: (1) Facial expression obtained from a capsule network and (2) Facial attributes trained based on VGG-face -Scene: scene features extracted using a capsule network from the whole image
(Correa et al., 2018)	AMIGOS	Individual-level affect analysis	-EEG -ECG -GSR	-	Various features extracted from EEG, ECG and GSR signals

features were utilised to predict group-level affect information in (Dhall, Goecke and Gedeon, 2015).

For group-level affect analysis, a multi-modal framework can be used to combine bottom-up and top-down approaches. For example, Dhall *et al.* (Dhall, Joshi, Sikka, Goecke and Sebe, 2015) used facial features, i.e., the facial action unit and low-level facial features, as the bottom-up component, and they considered scene information as the top-down component. Similar works that combine both face-centred information and image-based context/scene information have also been conducted in (Li et al., 2016; Tan et al., 2017). In addition to face and context information, Mou *et al.* (Mou et al., 2015) also used body features to predict the valence and arousal of a group of people, but this method is limited to experiment on specific groups based on the fixed number of faces and bodies. A similar work was conducted by Huang *et al.* (Huang et al., 2018), where a multi-modal framework for group affect prediction using face, upper body and scene information was introduced and an information aggregation method was proposed for generating feature descriptions of face, upper body, and scene. At a further step, the winner of EmotiW 2018 (Dhall et al., 2018) (Guo et al., 2018) introduced a hybrid framework for the recognition of group-level emotion in an image, which include more cues, i.e., faces, scenes, skeletons and salient regions, and features that are extracted based on deep neural networks, and finally all of them are combined to classify the emotions. Miranda-Correa *et al.* (Correa et al., 2018) integrated different neuro-physiological signals, i.e., EEG, ECG and GSR, for personality and individual-level affect recognition in both group and individual settings.

Even though a lot of works have conducted multi-modal affect analysis and show that multi-modal frameworks outperform uni-modal frameworks, it still needs to be further investigated to encode more complementary information among different modalities, while removing redundant information. In addition, when it comes to group settings, we know that emotion expressed by a person can impact other persons in the scene or in the interaction, therefore it is also important to find a way to model the inter personal emotion dependency in a multi-modal system.

3.2.3 Analysis of other attributes in group settings

In addition to affect, a number of works focus on group analysis from some other perspectives, such as emergent leader recognition (Sanchez-Cortes et al., 2012), dominant person detection (Aran and Gatica-Perez, 2010), social role recognition (Zancanaro et al., 2006; Pianesi et al., 2007; Sapru and Bourlard, 2015), group cohesion analysis (Hung and Gatica-Perez, 2010), and group satisfaction (Lai and Murray, 2018), which are closely connected to the affect analysis in group settings.

Gallagher and Chen (Gallagher and Chen, 2009) proposed a method to recognise the age and gender of individuals in group images using a type of contextual features, the structure and distribution of people in group images. In (Zhu et al., 2014), the authors introduced a framework to analyse the the Quality of the Viewing Experience in terms of five aspects, namely enjoyment, endurability, satisfaction, involvement in the viewing experience and perceived visual quality in both individual and group settings. Sanchez-Cortes *et al.* (Sanchez-Cortes et al., 2012) presented a computational framework to identify emergent leaders in small groups using nonverbal behaviours, where a new Emergent LEAdership (ELEA) dataset was collected and annotated. ELEA consists of 40 recorded videos with a group of people (3 or 4) discuss a hypothetical survival, which has also been used for group decision-making performance (Avci and Aran, 2016) and personality analysis (Aran and Gatica-Perez, 2013). Avci *et al.* (Avci and Aran, 2014) studied the relationship of a group's performance with the interaction between group members and the individuals' personality traits using the audio and visual nonverbal behaviours. Hung *et al.* (Hung and Gatica-Perez, 2010) investigated group cohesion estimation by utilising audio, visual, and audio-visual cues, such as activity of each person and motion information. Leite *et al.* (Leite et al., 2015) studied the individual engagement estimation in group settings in the context of human-robot interaction. In addition, some previous works on group-level analysis focused on group activity recognition (Lan, Sigal and Mori, 2012; Lan, Wang, Yang, Robinovitch and Mori, 2012). Most of the aforementioned works analyse what is happening within the group. Only recently, works focusing on automatic analysis of the relationship between the members of different groups have emerged. Correa *et al.* (Mioranda-Correa and Patras, 2018) predicted the social context, i.e., whether a person is alone or in a group, using neuro-physiological signals. In this thesis, we present a novel framework for social context prediction of (1) whether a person is being alone or in a group and (2) group membership recognition, i.e., recognition of which group each individual belongs to, using non-verbal behaviours. In Chapter 4 and Chapter 6 we will introduce the frameworks for the prediction of whether a person is alone or in a group, and the recognition of group membership respectively.

3.3 Summary

In this chapter, we review the related works in affect analysis and stress the differences and challenges for the affect analysis in group settings.

From static to dynamic. Emotion is a continuous process rather than a static status, therefore, it is obvious that the affects of individuals can be better analysed by investigating the cues demonstrated in dynamic videos but this also brings challenges due to the complex dynamics of affective states in videos. To encode the spatio-temporal information in dynamic videos, temporal models either using the hand-crafted features or deep learning based approaches can be employed. For the hand-crafted features, the static features from single image can be extended to represent a sequence of frames using Three Orthogonal Planes, such as extending LBP to LBP-TOP, or dense trajectories can be used to extract the temporal information. For deep learning based methods, 3D convolution operations can be used instead of 2D ones to extract representations for a video. The temporal modelling method RNN is also popular for encoding temporal information and has shown good performance for affect analysis.

From individual to group settings. Affect analysis has moved from individual settings, i.e., one person in an image or in a video, to group settings, where there are multiple people in a scene. Compared to the databases for affect analysis in individual settings, there are a small number of databases for affect analysis in group settings. HAPPEI (Dhall, Goecke and Gedeon, 2015), GAF (Dhall, Joshi, Sikka, Goecke and Sebe, 2015) and MultiEmoVA (Mou et al., 2015) are all static images and labelled with only a small number of categorical emotions. AMIGOS consists of dynamic videos and is labelled with continuous valence and arousal dimensions, but limited to the scenario of watching multimedia content. To further advance the investigation into affect analysis in group settings, it is important to collect and annotate more diverse datasets in different scenarios.

From uni-modal to multi-modal analysis. Face has been the most prominent cue for affect analysis, both in individual settings and group settings. The facial features can be divided into appearance and geometric representations to represent the texture information, e.g., wrinkles and bulges on the faces, and the shape of different parts of faces, e.g., eyes and mouth, respectively. They are both important and complimentary to each other. As humans express and understand emotions using multiple channels, including facial expressions, hand and body gestures etc., it is also important to utilise features out of other cues in addition to facial information, especially for group settings. In group settings, there are always multiple people appearing in a scene or in an interaction, the face of one person may be occluded by the other person or may be far from the camera due to the motion, when body information is important for affect analysis. In addition, context is also important for affect analysis, whom a person is staying with, where she/he is and what she/he is doing. Therefore, it is important to explore how to do affect analysis efficiently in a multi-modal manner especially in group settings.

Chapter 4

AFFECT RECOGNITION IN INDIVIDUAL AND GROUP VIDEOS

4.1 Introduction

Over the last decades, various methodologies have been proposed to automate the analysis of affect and emotions. The majority of the existing works focus on individual settings and so far little attention has been paid to the analysis in group settings, either at the overall group-level emotion displayed by the entire group or at the individual-level emotion displayed by each individual within that group. However, in the real world, people are very often with others, interacting in group settings, such as in a meeting and in a party. To this end, it is important to study the affect expressed by people in group settings.

In this chapter, we focus on affect recognition in both individual and group settings, which paves the way for the further analysis in the following chapters. In details, we aim to investigate the following questions: (1) whether the method used for affect recognition of individuals in individual settings can be transferred to group settings. That is whether it is possible to recognise the affect expressed by each individual while presented with movie stimuli in both individual and group settings using body and facial features; (2) making use of the affective behavioural cues exhibited in different settings, whether it is possible to predict a person is in an individual setting or in a group setting.

To investigate the above mentioned questions, we introduce a framework for affect recognition using different facial and body features for both individual and group settings. A set of experiments is carried out on databases containing both individual and group videos. From the experiments, we (1) find the method used for affect recognition of a person in individual settings (facial or body behaviours with classifier or regressor) can be transferred to the affect recognition of individuals in group settings for AMIGOS dataset; (2) confirm body information can support affect recognition not only in individual settings but also in group settings; (3) find that facial and body behaviours can be used to predict whether a person is in an individual setting or in a group setting, i.e., being alone or in a group.

The remaining part of this chapter is structured as follows: the proposed methods for affect and context recognition in individual and group videos are stated in Section 4.2; the databases, both individual and group videos, are introduced in Subsection 2.4.3; the experimental results and analysis are presented and discussed in Section 4.4; and finally in Section 4.5 conclusion and future work are presented.

4.2 The proposed framework

We propose frameworks to recognise (1) the affect of individuals in different settings, i.e., individual and group videos and (2) the prediction of contextual information, i.e., whether a person is in an individual setting or in a group setting by using non-verbal behavioural cues, i.e., face and body cues.

We first adopt an SVM-based multi-modal method using dynamic features and conduct experiments on both individual and group videos. The proposed framework is illustrated in Figure 4.1. To represent faces, we use geometric and appearance representations. The geometric feature we utilize is facial landmark trajectory, while appearance feature we use is

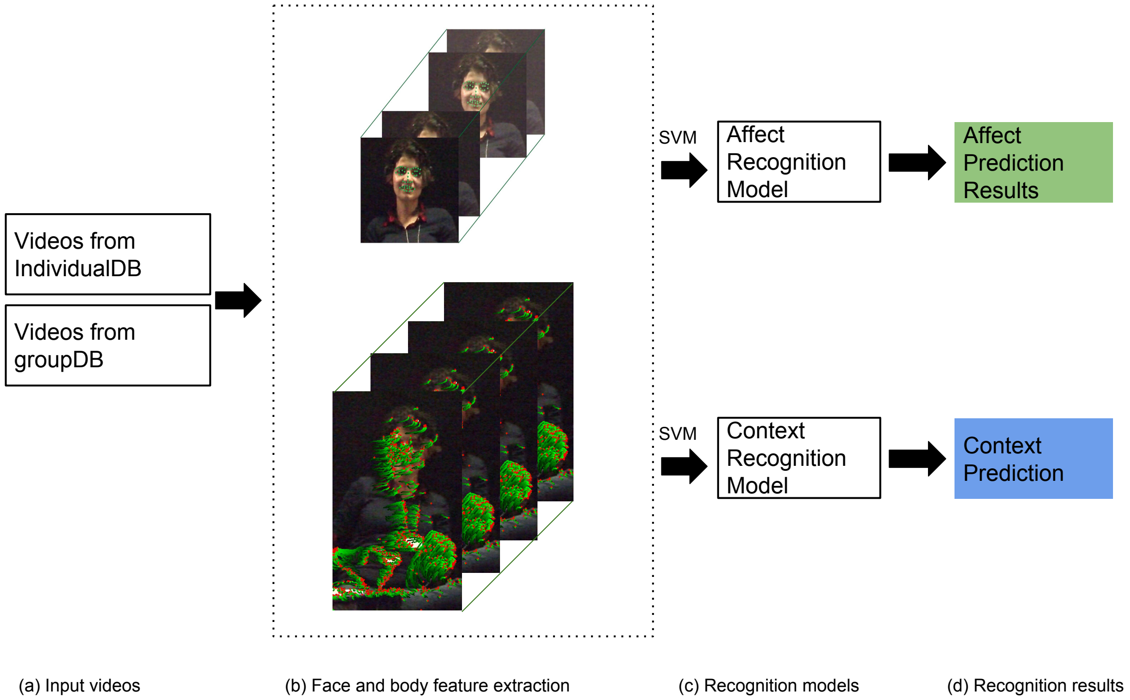


Figure 4.1 Description of the proposed framework. (a) Input videos, individual videos from IndividualDB and group videos from GroupDB, the details of the databases are provided in Subsection 2.4.3; (b) Feature extraction - face and body features extracted; (c) Different recognition models are trained, i.e., affect recognition models and contextual information recognition models.

the extended volume Quantised Local Zernike Moments (QLZM) (Sariyanidi et al., 2013; 2015) extracted along facial landmark trajectories. In light of the body representations, we first extract dense trajectories and then we extract Histogram of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF) descriptors along each trajectory (Wang et al., 2013). Before feeding the features to different classifiers and regressors, we encode the different face and body low-level descriptors into Fisher Vectors (FV). Multiple experiments are carried out for affect analysis using unimodal and multi-modal cues.

In a further step, we use a temporal learning model, Long Short-Term Memory (LSTM) Networks (Hochreiter and Schmidhuber, 1997) for affect recognition. LSTM is one of the state-of-the-art sequence modeling approaches and has been successfully applied to affect analysis (Li et al., 2016; Chen et al., 2017).

Details of LSTM. Long Short Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) is a special type of Recurrent Neural Networks (RNNs). LSTMs has repeating modules of neural network in the form of a chain as shown in Figure 4.2. Each of the module is also called an LSTM unit, which is composed of a cell state, an input gate, a forget gate and an output gate.

Here we will explain the cell and all of the gates using the example shown in Figure 4.2. Firstly, LSTM can decide what information to forget/throw away from the cell state (c_{t-1}) using the "forget gate" achieved by a sigmoid layer. This sigmoid layer takes h_{t-1} and x_t as input and outputs f_t as shown in Equation 4.1.

$$f_t = \sigma(W_f h_{t-1} + W_f x_t + b_f) \tag{4.1}$$

Secondly, an "input gate" and a $\tilde{c}_t$ will decide what new information will be stored into the cell state. This is achieved using a sigmoid layer and a tanh layer as shown in Equation 4.2 and Equation 4.3. The sigmoid layer takes h_{t-1} and x_t as input and outputs i_t. The tanh layer also takes h_{t-1} and x_t as inputs and generates a vector of new candidate values, i.e., $\tilde{c}_t$.

$$i_t = \sigma(W_i h_{t-1} + W_i x_t + b_i) \tag{4.2}$$

$$\tilde{c}_t = tanh(W_c h_{t-1} + W_c x_t + b_c) \tag{4.3}$$

The next step is to update the cell state as shown in Equation 4.4. i_t decides how much of $\tilde{c}_t$ will be used and f_t from Equation 4.1 decides how much to forget the earlier state.

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{4.4}$$

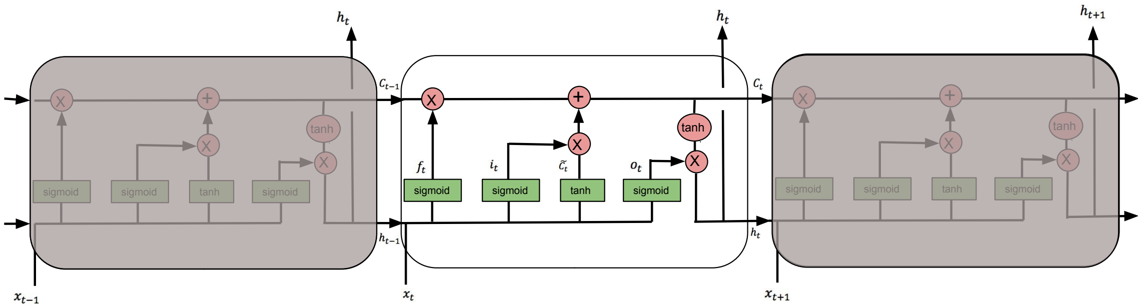


Figure 4.2 Illustration of the repeating modules in an LSTM containing *sigmoid* layers and one *tanh* layer. The green rectangle represents neural network layer; the red circle represents point-wise plus or multiplication; an arrow refers to the vector transfer; arrows getting together denotes the concatenation of the vectors; and a line forking refers to the copy of the content.

Finally, it is the output, which is based on the cell state, c_t, and the output gate, o_t. The output gate is with a sigmoid layer and takes h_{t-1} and x_t as inputs as shown in Equation 4.5. Then, the cell state, c_t, passes a tanh layer and multiply o_t to get the output h_t as shown in Equation 4.6.

$$o_t = \sigma(W_o h_{t-1} + W_o x_t + b_o) \tag{4.5}$$

$$h_t = o_t * tanh(c_t) \tag{4.6}$$

In this manner, a cell state, an input gate, a forget gate and an output gate together decide the output and keep the parts that are needed. Note that all of the W and b from Equation 4.1 to Equation 4.6 denote weights and bias respectively.

The framework with LSTM used for affect recognition is illustrated in Figure 4.3. Frame-level static features are first extracted from the input sequences. The static features extracted from each frame are then fed into a one-layer LSTM. After fully connected layers, the affect recognition results are obtained.

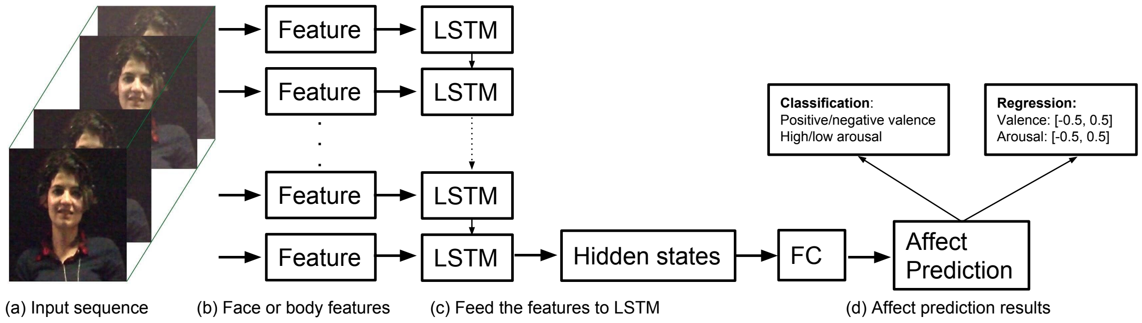


Figure 4.3 Illustration of the approach for affect analysis using LSTM. (a) Input sequence. (b) Frame-level features are extracted. (c) Features extracted from every frame are fed into a one-layer LSTM with 128 hidden states. (d) Affect prediction results obtained for either classification or regression.

4.3 Feature extraction for affect analysis

4.3.1 Face features

Before extracting facial features, we first utilise Intraface (Xiong and De la Torre, 2013) to detect facial landmarks of each face in the video. After applying Intraface, each face obtains 49 facial points. However, not all faces are detected due to illumination, occlusion, and pose variations in such a naturalistic scenario. In order to make the facial feature extraction consistent among all frames, when the face detection fails in a current frame, the position of the last detected face is used.

In terms of facial geometric features, let $X_t = [\,(x_t^1,\, y_t^1),\, (x_t^2,\, y_t^2)\, ...\, (x_t^n,\, y_t^n)\,]$ denote the positions of n landmark points of the face at the current frame t. The number of landmark points on each face $n = 49$. x_t^k and y_t^k refer to the coordinates of the k-th landmark point at the current frame t. Then landmark points of the subsequent frames are concatenated to generate the facial landmark trajectories. In this way, the representation of the facial landmark trajectory encodes the motion patterns of the facial points as the body trajectories used in (Wang et al., 2013). The k-th facial landmark point is described by a sequence $(\Delta X_t^k,\, \Delta X_{t+1}^k\, ...\, \Delta X_{t+L-1}^k)$ of displacement vectors, where $\Delta X_t^k = (X_{t+1}^k - X_t^k) = (x_{t+1}^k -$

$x_t^k, y_{t+1}^k - y_t^k$) and L is the length of the facial landmark trajectories. The obtained vector is then normalised by the sum of the displacement vector magnitudes:

$$Y^k = \frac{(\Delta X_t^k, \Delta X_{t+1}^k ... \Delta X_{t+L-1}^k)}{\sum\limits_{j=t}^{t+L-1} ||\Delta X_j^k||}. \tag{4.7}$$

Y^k is referred as *Facial Landmarks* in the remaining part of this chapter. The length of the facial landmark trajectories is fixed as $L = 15$ frames based on (Wang et al., 2013). In this way, a 30 (30 = $2 \times L$, where $L = 15$) dimensional feature is generated around each landmark point of the face. And for each face, the dimensionality of the descriptor is 49×30 as 49 landmark points are detected for each face.

After extracting the geometric features, Quantised Local Zernike Moments (QLZM) (Sariyanidi et al., 2013) obtained from the local patch around each facial landmark point are extracted as the facial appearance representation. QLZM is one of the state-of-art facial features for affect recognition and were originally designed for static images (Sariyanidi et al., 2013).

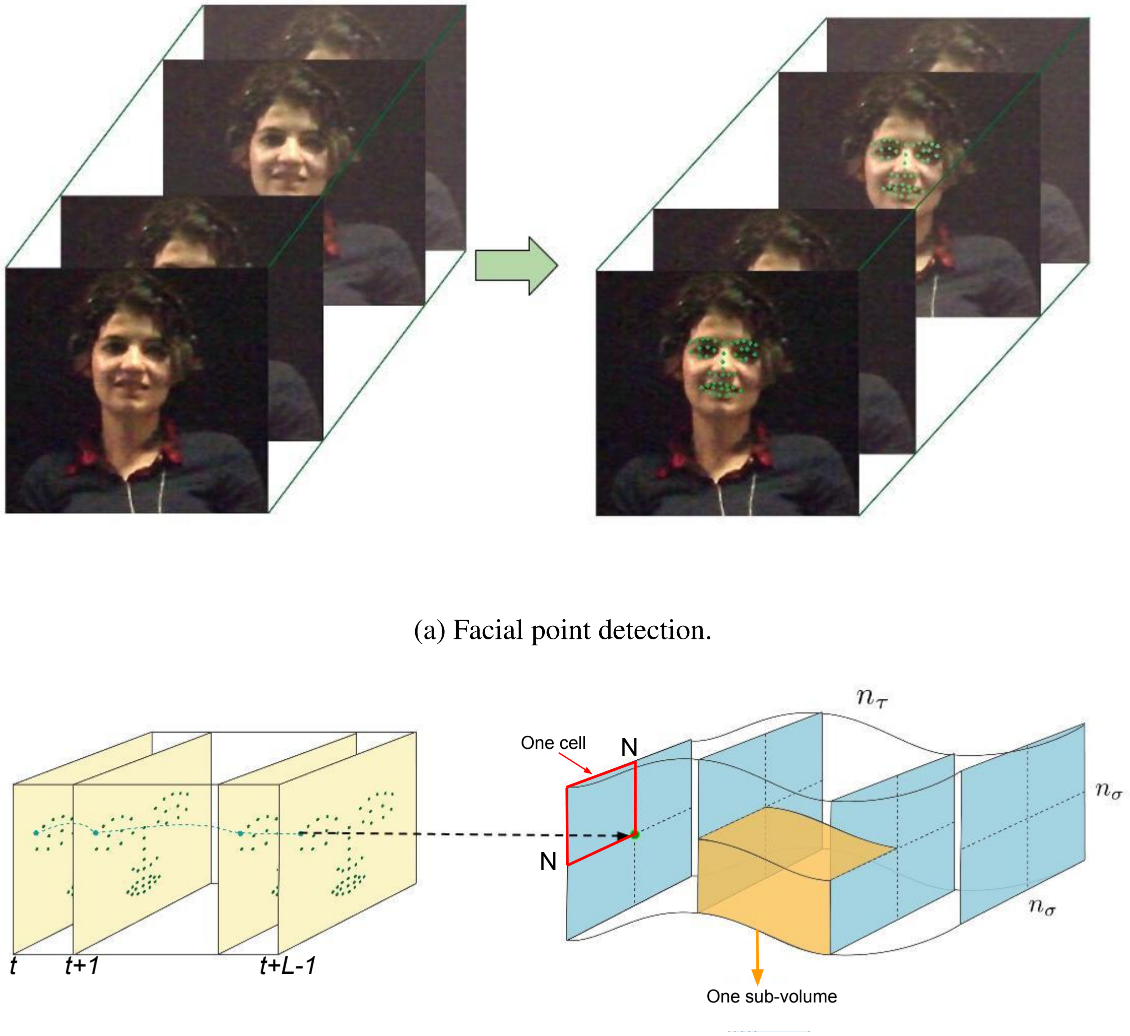
However, as we are focusing on video information processing, temporal information is important. Therefore, it is extended to a volume representation to embed both spatial and temporal information, as described in Figure 4.4. We refer to the facial appearance feature, volume based Quantised Local Zernike Moments, as *vQLZM* in the remaining part of the thesis. The size of the volume is $N \times N$ pixels, while the length is $L = 15$ frames, the same volume size with the *Facial Landmarks*. The volume is then subdivided into a $n_\tau \times n_\sigma \times n_\sigma$ sub-volumes. To get the vQLZM features, firstly the static QLZM descriptor is computed in each cell, which is in size of N/n_σ as shown in Figure 4.4b. The dimension of the descriptor is 4. Secondly, the descriptors in a sub-volume are averaged to be one descriptor, which is still in a dimension of 4. Then the final descriptor is generated by concatenating these descriptors of each sub-volume, therefore, the final descriptor of each volume is $4 \times n_\tau \times n_\sigma \times n_\sigma$. $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$ are used, which is the same as (Wang et al., 2013). Thus the final dimension of the descriptor of each volume is $4 \times 3 \times 2 \times 2 = 48$.

4.3.2 Body features

Body feature extraction is a type of person-based representation, therefore, the first step is to apply a person detector. Constrained by our experimental setups - a fixed number of people in the video (either one in individual videos or four in group videos) and a static camera, we use an ad-hoc scheme that is to use only the central part where the person is in individual videos and equally divide the frame in four parts in group videos. In order to avoid the overlap between the participants that are neighbouring each other, we leave a space between every two neighbours. The space size is equal to the average size of the faces across all videos, i.e., 64. Then, dense trajectories (Wang et al., 2013) are extracted. Trajectories capture the local motion information of the video and dense representation guarantees a good coverage of foreground motion as well as of the surrounding context. Subsequently, HOG and HOF features are obtained along each extracted trajectory. They are computed in the spatio-temporal volume aligned with the trajectories as shown in Figure 4.5. HOG and HOF orientations are quantized into eight bins with full orientations. However, as an additional zero bin is added for HOF for pixels with optical flow magnitudes lower than the threshold (i.e., nine bins in total), the final representation size of HOG is 96 and that of HOF is 108 with the trajectory length $L = 15$ frames. We refer to these two body related representations as *body HOG* and *body HOF* respectively in the rest of the thesis.

4.3.3 Fisher vector encoding

Fisher Vector (FV) representation (Sánchez et al., 2013) has been widely utilised in traditional computer vision problems (e.g., action recognition (Wang et al., 2013; Wang and Schmid, 2013)) and affect analysis (e.g., depression analysis (Jain et al., 2014; Dhall and Goecke,



(a) Facial point detection.

(b) Volume QLZM extraction.

Figure 4.4 Details of the approach to extract the *vQLZM* facial appearance feature. Figure (a) shows the detection of facial landmark points. Figure (b) illustrates the tracking of facial landmark point over L frames. QLZM is extracted over a local neighbourhood of $N \times N$ pixels along each landmark point. To encode the structure information, the local volume is subdivided into $n_\tau \times n_\sigma \times n_\sigma$ sub-volumes. $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$ as used in (Wang et al., 2013).

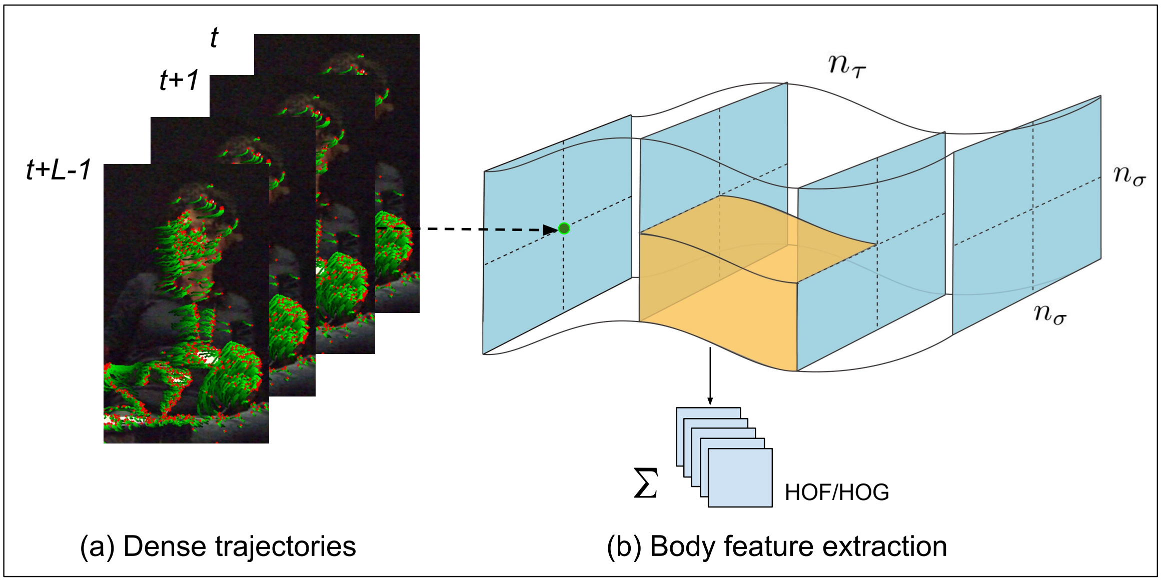


Figure 4.5 Description of the method of body HOG/HOF feature extraction. (a) shows the detected dense trajectories. (b) illustrates the HOG/HOF feature extraction along the trajectories in the spatial scale over L frames. Motion information over a local neighbourhood of $N \times N$ pixels along each trajectory point are extracted. In order to encode the structure information, the local volume is divided into small spatio-temporal grid of size $n_\tau \times n_\sigma$. Based on (Wang et al., 2013), $n_\tau = 3$, $n_\sigma = 2$ and $L = 15$.

2015)). The first work that applied Fisher Vector descriptors for the problem of action recognition in videos used local features extracted along dense trajectories (Wang et al., 2011). The trajectories are extracted by defining a dense grid of points which are then tracked using optical flow that was estimated offline to include motion information in the pipeline. By encoding the extracted trajectory features with the Fisher Vector descriptor, this approach and its improved version (Wang et al., 2013; Wang and Schmid, 2013) achieved the state-of-the-art results for the action recognition before deep neural networks are widely utilised. It encodes both the first and second order statistics between the low-level (local) video/image descriptors and a Gaussian Mixture Model (GMM). To obtain the Fisher Vector, firstly, Principal Component Analysis (PCA) is applied to the descriptors to decrease the dimensionality. Secondly, the low-level descriptors (i.e., face and body descriptors in our case) is fitted to a GMM. The covariance matrices used for GMM here are diagonal. As suggested by (Wang et al., 2013; Wang and Schmid, 2013), the number of Gaussians is set to $K = 256$ and randomly selected 256,000 descriptors are used to fit a GMM. The dimensionality of the Fisher Vector is $(2D+1)K$ (D refers to the dimensionality of the descriptor before feeding to GMM, i.e., after applying PCA), which is used to represent one clip. Four different types of Fisher Vectors (FVs) are generated based on face and body features, namely, *Facial Landmarks-FV, vQLZM-FV, body HOG-FV* and *body HOF-FV*. Dynamic features utilized in this thesis refer to these four features, i.e., *Facial Landmarks-FV, vQLZM-FV, body HOG-FV* and *body HOF-FV*, while the static features refer to features extracted from static frames, e.g., QLZM extracted from each frame.

4.4 Experiments and discussions

The experiments are carried out using IndividualDB and GroupDB, two databases for studying affect analysis from multi-modal cues in different settings, i.e., individual settings and group settings respectively.

4.4.1 Implementation details

Data

For GroupDB, group videos from four groups are used in the experiments, i.e., three groups (twelve subjects) with recordings of people watching four movies (N1, P1, B1 and U1) and one group (four subjects) with recordings of people watching three movies (B1, N1 and U1). In this case, we have data from sixteen subjects and fifteen sessions in total used in the experiments. One session refers to the recording of one group watching one movie. For each session, 20-second clips in line with the annotations labelled are utilised. The number of the 20-second clips from different sessions varies with the length of the movies, i.e., 70 clips for N1, 70 clips for B1, 56 clips for P1 and 42 clips for U1. As a result, the total number of clips we use in our experiments is $(70(B1) \times 4(4subjects) \times 4(4groups)) + (70(N1) \times 4(4subjects) \times 4(4groups)) + (56(P1) \times 4(4subjects) \times 3(3groups)) + (42(U1) \times 4(4subjects) \times 4(4gr-oups)) = 3,584$. In terms of IndividualDB, videos from 17 participants are used in the experiments. Each participant was recorded while watching 4 movies (N1, P1, B1 and U1). We also use 20-second clips. Therefore, the total number of clips we use in the experiments is $(70+70+56+42) \times 17 = 4,046$.

Experimental setup

Classification and regression models are built with different cross-validation setups, such as *subject-specific* and *leave-one-subject-out*. The parameters of each model are optimized over the training-validation data. *Subject-specific* refers to training the model using *leave-one-sample-out* cross-validation for the data of each subject separately. Namely, in each fold, one sample from a certain subject is used as testing data and all the other samples from the same subject are used as training data. In order to avoid the subject-dependency problem caused by the *subject-specific* model, *leave-one-subject-out* cross-validation is also applied. *Leave-one-subject-out* means that we use one subject's data for testing and all other subjects' data for training-validation in each fold. For GroupDB, *leave-one-group-out* cross-validation is also applied. *Leave-one-group-out* validation means that we use data from three groups out of four groups as training data, and data from the left one group as the testing data. For affect analysis, we did both classification and regression. Classification is formulated as a binary classification problem by quantizing both arousal and valence annotations into two classes using the median of all of the annotations as thresholds. In this way, arousal is quantized into *high* and *low* arousal and valence is quantized into *positive* and *negative* valence. The distribution of samples for GroupDB and IndividualDB along both arousal and valence dimensions after quantization is shown in Table 4.2. For contextual information prediction, it is formulated as a binary classification problem. We conduct experiments to

Table 4.1 The dimensionality of different features.

	Raw features extracted along the trajectories	After PCA	Fisher Vectors	PCA
	GroupDB			
QLZM	48	30	15616	672
Landmarks	30	26	13568	2542
HOG3D	48	17	8960	608
HOG-face	96	34	17664	36
HOG-body	96	48	24832	115
HOF-body	108	54	27604	870
	IndividualDB			
QLZM	48	30	15616	44
Landmarks	30	26	13568	1777
HOG-body	96	48	24832	234
HOF-body	108	54	27904	1567

Table 4.2 The distribution of samples for IndividualDB and GroupDB after quantizing the continuous annotations into two classes along both arousal and valence dimensions. Using the median of all of the annotations as thresholds, arousal is quantized into *high* and *low* arousal and valence is quantized into *positive* and *negative* valence.

Dimensions	Arousal		Valence	
Labels	High	Low	Positive	Negative
GroupDB	1,792	1,792	1,792	1,792
IndividualDB	2,023	2,023	2,023	2,023

predict whether a person is being alone or in a group based on face and body behavioural cues using *leave-one-subject-out* cross-validation.

Classifier used for affect recognition

In the first session of affect analysis, we conduct experiments using Support Vector Machines (SVM) (Chang and Lin, 2011) for classification and Support Vector Regression (SVR) for regression, with all extracted face and body features. In the second step of affect analysis, we conduct experiments on affect analysis using Long Short-Term Memory (LSTM) Networks (Hochreiter and Schmidhuber, 1997) with the best performing feature obtained from the first session.

Affect analysis is divided into two parts, i.e., affect classification and regression along arousal and valence dimensions. The first part of the experiments is affect recognition that is conducted using (1) different unimodal cues and (2) decision-level fusion of four different features, i.e., vQLZM-FV, Facial Landmarks-FV, body HOG-FV and body HOF-FV. As we use SVM as the classifier, decision-fusion is applied on the soft outputs of the single-modality classifiers. We utilise the publicly available SVM library LibSVM (Chang and Lin, 2011) for training and testing. Before the face and body features are fed to any classifier or regressor, we first apply PCA to reduce the dimensionality by preserving 99% of the variance. The second part of the affect recognition is carried out on the best performed unimodal feature QLZM using LSTM implemented on PyTorch platform (Paszke et al., 2017).

4.4.2 Experimental results and analysis

In this section, the affect recognition results are provided and discussed based on the two databases, IndividualDB and GroupDB separately. In addition, the context recognition results are reported in terms of prediction of whether a person is in an individual setting or in a group setting, i.e., alone or in a group.

Affect recognition in group settings

We utilised linear Support Vector Machine (SVM) to do classification and regression w.r.t. the dimensions along arousal and valence. The classification results obtained using unimodal features and decision-level fusion are illustrated in Table 4.3. Firstly, we can see that different types of features perform differently. Generally, vQLZM shows the best performance in both *leave-one-subject-out* and *subject-specific* cross-validation. It indicates that the proposed vQLZM descriptors are informative for tasks of affect analysis. Secondly, to compare with other features, one of representative handcrafted spatio-temporal features, HOG 3D (Klaser et al., 2008; Jung et al., 2015; Yang, Ciftci and Yin, 2018) is used. As listed in Table 4.3, the classification results obtained using HOG 3D with *leave-one-subject-out* setup are 0.64 along arousal and 0.60 along valence in terms of F1 score, which are not as good as that obtained using vQLZM (0.68 for arousal and 0.68 for valence). Thirdly, generally we can see that face features performing better than body features, for example, HOG features are used both for face and body, i.e., face HOG and body HOG respectively. From results shown in Table 4.3, we can see that face HOG performs slightly better than body HOG. For instance, in *leave-one-subject-out* setup, 0.61 and 0.60 are obtained for face HOG in terms of F1 score along arousal and valence respectively, while that for body are 0.57 and 0.58. However, we can confirm that body features individually are capable of recognising affect in group settings. Therefore, while facial information is lost in group settings due to occlusion etc., body information can be used for affect recognition. In addition, we can see that compared to *leave-one-subject-out* models, *subject-specific* models perform better due to the subject-dependency. The best results obtained in *leave-one-subject-out* setup are 0.69 along arousal and 0.68 along valence in terms of F1 score, while the best results obtained in *subject-specific* setup are 0.80 along arousal and 0.80 along valence. Therefore, if we have enough data, it will be good to have a specific model for affect recognition for each person as the person specific model is more accurate.

In terms of decision-level fusion, the decision values that are the obtained probabilities for all classes from individual features are given as input to an SVM. The results show that the classification performance using decision-fusion of four face and body features is generally better than or equal to the best results obtained with unimodal features for most of the time. In *leave-one-subject-out* setup, the best affect classification results obtained using unimodal cues are 0.68 along arousal and 0.68 along valence using the $F1$ score as our evaluation method; and those classification results of affect analysis obtained using decision fusion are 0.69 in terms of arousal and 0.68 in terms of valence. In *subject-specific* setup, the best affect classification results obtained using unimodal cues and decision fusion are the same, i.e., 0.80 in terms of arousal and 0.80 in terms of valence using the $F1$ score as our evaluation method.

For the regression of the affect analysis, we utilise Support Vector Regression (SVR) with a radial basis function (RBF) kernel. The results obtained with unimodal and multi-modal features are presented in Table 4.5. For the unimodal results, we can see that the regression results are quite similar to the classification ones, i.e., *vQLZM* generally performs best among all unimodal features. As to the decision-level fusion, we proceed in a similar way to the fusion in affect classification. Specifically, we fuse the ratings predicted from unimodal features in an SVR. In most of the cases, the performance of decision-fusion is better than or equal to the performance of the unimodal features. For example, in *subject-specific* setup, the best results in terms of CCC obtained using unimodal cues are 0.54 along arousal and 0.56 along valence, while those results obtained using decision-level fusion of face and body cues are 0.57 along arousal and 0.62 along valence. In *leave-one-subject-out* setup, the best results in terms of CCC obtained using unimodal cues and decision fusion are the same along arousal, i.e., 0.44.

Subsequently, we utilize LSTM and facial QLZM feature for affect classification and regression. LSTM is one of the state-of-the-art temporal modelling methods and facial QLZM feature is the best performed unimodal representation as shown in Table 4.3 and 4.5. The classification and regression results are reported in Table 4.4 and 4.6 respectively. We can see that compared to *vQLZM with SVM* and even *the decision-fusion with SVM*, LSTM is more powerful for arousal and valence recognition in dynamic videos. For example, the best regression results obtained with *dynamic features with SVR* are 0.44 along arousal and 0.53 along valence in terms of CCC, while these obtained with *LSTM* are 0.65 along arousal and 0.70 along valence. To show the results clearly, we compare the classification and regression results obtained with *vQLZM with SVM*, *decision-fusion with SVM* and *QLZM with LSTM*

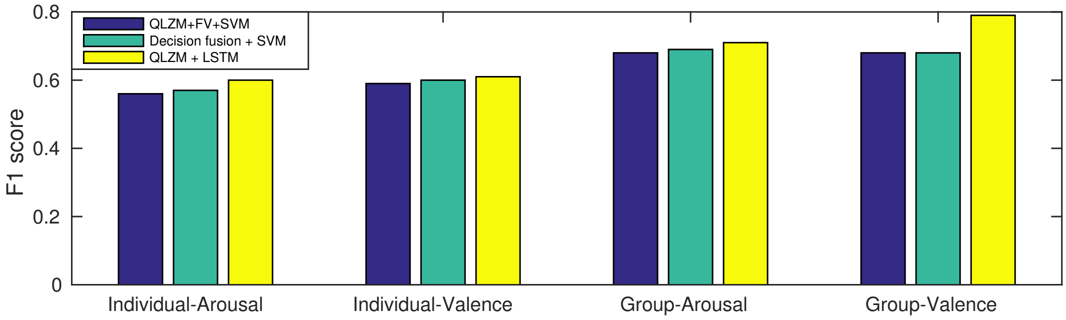


Figure 4.6 Illustration of the affect classification results in terms of F1 score using different features and classifiers for individual and group settings along arousal and valence dimensions in *leave-one-subject-out* setup.

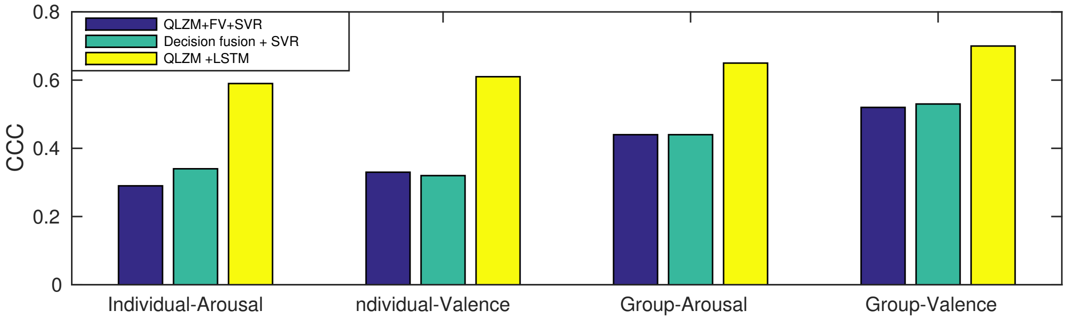


Figure 4.7 Illustration of the affect regression results in terms of CCC using different features and regression methods for individual and group settings along arousal and valence dimensions in *leave-one-subject-out* setup.

in Figure 4.6 and 4.7. The results clearly show that LSTM improves affect recognition performance as has previously been reported in single-person videos in (Chen et al., 2017).

Affect recognition in individual settings

Similar to affect recognition in GroupDB, Support Vector Machine (SVM) is utilised to do classification and regression w.r.t. the dimensions along arousal and valence. The classification and regression results using four different unimodal features and decision-level fusion are illustrated in Table 4.7 and Table 4.9 respectively. It can be seen that the results are consistent with the results obtained using GroupDB: (1) different features provide different classification/regression results and *vQLZM* outperforms all the other unimodal features in

Table 4.3 The affect classification results in terms of F1 score for **GroupDB** using **SVM** with unimodal face and body features and the decision-level fusion of vQLZM-FV, Facial Landmarks-FV, body HOG-FV and body HOF-FV. The standard deviation (std) is also presented (bold for the best results).

Dimensions	Arousal	Valence
	$F_1(std)$	$F_1(std)$
Chance level	0.5	0.5
Leave-one-subject-out		
Face		
vQLZM-FV	0.68 (0.12)	**0.68 (0.14)**
HOG3D-FV	0.64 (0.12)	0.60 (0.11)
face HOG-FV	0.61 (0.09)	0.60 (0.12)
Landmarks-FV	0.61 (0.08)	0.58 (0.07)
Body		
body TRA-FV	0.55 (0.09)	0.60 (0.07)
body HOG-FV	0.57 (0.12)	0.58 (0.07)
body HOF-FV	0.61 (0.07)	0.59 (0.08)
Decision-fusion of four features	***0.69 (0.14)***	***0.68 (0.15)***
Subject-specific		
Face		
vQLZM-FV	**0.80 (0.07)**	**0.80 (0.06)**
HOG 3D-FV	0.79 (0.09)	0.79 (0.06)
face HOG-FV	0.76 (0.12)	0.71 (0.11)
Landmarks-FV	0.69 (0.13)	0.64 (0.09)
Body		
body TRA-FV	0.70 (0.11)	0.69 (0.10)
body HOG-FV	0.70 (0.12)	0.68 (0.11)
body HOF-FV	0.70 (0.12)	0.67 (0.13)
Decision-fusion of four features	***0.80 (0.07)***	***0.80 (0.06)***

Table 4.4 The affect classification results in terms of F1 score for **GroupDB** with static **QLZM** features using **LSTM**. The standard deviation (std) is also presented in parentheses.

Dimensions	Arousal	Valence
	$F_1(std)$	$F_1(std)$
Chance level	0.5	0.5
Leave-one-subject-out		
GroupDB	**0.71 (0.09)**	**0.79 (0.11)**

Table 4.5 The affect regression results in terms of PCC and CCC for **GroupDB** using **SVR** with unimodal face and body features and the decision-level fusion. The standard deviations(std) are also reported (bold for the best results).

Dimensions	Arousal		Valence	
	PCC (std)	CCC (std)	PCC (std)	CCC (std)
Leave-one-subject-out				
Face				
vQLZM-FV	0.58 (0.08)	**0.44 (0.08)**	0.57 (0.13)	0.52 (0.07)
HOG 3D-FV	0.46 (0.13)	0.28 (0.07)	0.51 (0.12)	0.30 (0.06)
face HOG-FV	0.42 (0.16)	0.24 (0.07)	0.43 (0.16)	0.25 (.07)
Landmarks-FV	0.44 (0.12)	0.23 (0.06)	0.39 (0.17)	0.25 (0.04)
Body				
body TRA-FV	0.38 (0.18)	0.20 (0.04)	0.36 (0.19)	0.25 (0.06)
body HOG-FV	0.27 (0.25)	0.21 (0.05)	0.29 (0.22)	0.26 (0.06)
body HOF-FV	0.42 (0.20)	0.27 (0.05)	0.31 (0.18)	0.25 (0.05)
Decision-fusion of four features	***0.61 (0.08)***	***0.44 (0.10)***	***0.58 (0.12)***	***0.53 (0.08)***
Subject-specific				
Face				
vQLZM-FV	**0.71 (0.12)**	0.54 (0.08)	0.67 (0.16)	0.56 (0.17)
HOG 3D-FV	0.62 (0.12)	0.48 (0.08)	0.60 (0.19)	0.39 (0.17)
face HOG-FV	0.61 (0.11)	0.53 (0.08)	0.59 (0.16)	0.45 (0.18)
Landmarks-FV	0.54 (0.17)	0.30 (0.12)	0.46 (0.26)	0.32 (0.11)
Body				
body TRA-FV	0.52 (0.16)	0.27 (0.10)	0.45 (0.28)	0.31 (0.12)
body HOG-FV	0.58 (0.17)	0.45 (0.11)	0.56 (0.24)	0.47 (0.16)
body HOF-FV	0.53 (0.18)	0.32 (0.10)	0.48 (0.27)	0.35 (0.13)
Decision-fusion of four features	*0.70 (0.11)*	***0.57 (0.09)***	***0.69 (0.25)***	***0.62 (0.22)***

Table 4.6 The affect regression results in terms of PCC and CCC for **GroupDB** with static **QLZM** features using **LSTM**. The standard deviations(std) are also reported.

Dimensions	Arousal		Valence	
	PCC (std)	CCC (std)	PCC (std)	CCC (std)
Leave-one-subject-out				
GroupDB	**0.66 (0.09)**	**0.65 (0.10)**	**0.72 (0.11)**	**0.70 (0.15)**

Table 4.7 The affect classification results in terms of F1 score for **IndividualDB** using **SVM** with unimodal face and body features and the decision-level fusion. The standard deviation (std) is also presented in parentheses (bold for the best results).

Dimensions	Arousal	Valence
	$F_1(std)$	$F_1(std)$
Chance level	0.5	0.5
Leave-one-subject-out		
Face		
vQLZM-FV	0.56 (0.12)	0.59 (0.12)
Landmarks-FV	0.55 (0.07)	0.57 (0.07)
Body		
body HOG-FV	0.55 (0.12)	0.54 (0.09)
body HOF-FV	0.55 (0.07)	0.52 (0.07)
Decision-fusion of four features	***0.57 (0.13)***	***0.60 (0.11)***
Subject-specific		
Face		
vQLZM-FV	0.70 (0.15)	0.73 (0.12)
Landmarks-FV	0.66 (0.14)	0.66 (0.09)
Body		
body HOG-FV	0.70 (0.13)	0.69 (0.15)
body HOF-FV	0.66 (0.13)	0.66 (0.11)
Decision-fusion of four features	***0.72 (0.14)***	***0.75*** **(0.09)**

both classification and regression models; and (2) the results obtained with fusion of the facial and body features are generally better than or equal to those obtained with unimodal features. For example, in *subject-specific* setup, the best results in terms of CCC obtained using unimodal cues are 0.62 along arousal, while those results obtained using decision-level fusion of face and body cues are 0.66 along arousal. However, in *leave-one-subject-out* setup, the best result in terms of CCC obtained using unimodal cues for arousal is 0.29, while that for decision-fusion is 0.34. In addition, in Section 4.4.2 we propose a method that attempts to predict whether a person is in an individual setting or in a group setting using their non-verbal behaviours.

Table 4.8 The affect classification results in terms of F1 score for **IndividualDB** with static **QLZM** features using **LSTM**. The standard deviation (std) is also presented in parentheses.

Dimensions	Arousal	Valence
	$F_1(std)$	$F_1(std)$
Chance level	0.5	0.5
Leave-one-subject-out **IndividualDB**	**0.60 (0.12)**	**0.61 (0.15)**

Table 4.9 The affect regression results in terms of PCC and CCC for **IndividualDB** using **SVR** with unimodal face and body features and the decision-level fusion. The standard deviations (std) are also reported (bold for the best results).

Dimensions	Arousal		Valence	
	PCC (std)	CCC (std)	PCC (std)	CCC (std)
Leave-one-subject-out				
Face				
vQLZM-FV	0.34 (0.27)	0.29 (0.08)	0.34 (0.26)	**0.33 (0.08)**
Landmarks-FV	0.29 (0.22)	0.15 (0.04)	0.25 (0.20)	0.19 (0.05)
Body				
body HOG-FV	0.27 (0.25)	0.23 (0.05)	0.13 (0.11)	0.12 (0.03)
body HOF-FV	0.30 (0.22)	0.18 (0.04)	0.26 (0.21)	0.21 (0.05)
Decision-fusion of four features	***0.44 (0.29)***	***0.34 (0.09)***	***0.47 (0.25)***	*0.32 (0.09)*
Subject-specific				
Face				
vQLZM-FV	**0.76 (0.24)**	0.62 (0.17)	**0.69 (0.41)**	0.60 (0.23)
Facial Landmarks-FV	0.59 (0.25)	0.39 (0.12)	0.48 (0.34)	0.36 (0.20)
Body				
body HOG-FV	0.66 (0.29)	0.53 (0.19)	0.58 (0.38)	0.52 (0.19)
body HOF-FV	0.59 (0.24)	0.40 (0.13)	0.46 (0.34)	0.35 (0.14)
Decision-fusion of four features	*0.75 (0.33)*	***0.66 (0.19)***	***0.69 (0.41)***	***0.67 (0.24)***

Table 4.10 The affect regression results in terms of PCC and CCC for **IndividualDB** with static **QLZM** features using **LSTM**. The standard deviations(std) are also reported.

Dimensions	Arousal		Valence	
	PCC (std)	CCC (std)	PCC (std)	CCC (std)
Leave-one-subject-out **IndividualDB**	**0.60 (0.18)**	**0.59 (0.20)**	**0.62 (0.18)**	**0.61 (0.23)**

Table 4.11 The contextual recognition (whether a person is alone or in a group) results obtained with different face and body features. The table reports the average recognition accuracy over all subjects. The standard deviation of all subjects is also presented in parentheses. The chance level (50%) is also provided.

	Leave-one-subject-out
Chance level	50%
vQLZM	85% (32.0%)
Facial Landmarks	90% (24.5%)
body HOG	93% (23.2%)
body HOF	91% (23.6%)

Contextual information recognition

In addition, we investigate contextual information prediction using non-verbal behavioural cues. We conduct experiments to recognise whether a person is alone or in a group using the extracted face and body features described in Section 4.3. The results are shown in Table 4.11. We can see that the results we obtained, all above 85%, are significantly better than the chance level of 50%. In addition, it can be seen that body features perform slightly better than face features. It is possibly due to the fact that it is relatively difficult to utilise the facial information in this case as facial information is more subtle than body motion and gestures. Predicting whether a person is alone or in a group successfully indicates that people behave distinctly while they are alone compared to being within a group.

4.5 Discussion and conclusion

In this chapter, a framework is introduced for automatic affect recognition in both individual settings (i.e., IndividualDB) and group settings (i.e., GroupDB). To this end, different face and body features are extracted to analyse the affective states of individuals in terms of valence and arousal dimensions; and decision-level fusion was utilised to combine different facial and body features. A set of experiments on affect recognition is carried out on both individual and group videos and the results can be are concluded as follows. Firstly, the individual affect can be recognised using facial or body behaviours for both individual and group dataset; the proposed *vQLZM* descriptor outperforms the other unimodal features for the task of affect recognition for AMIGOS dateset. Secondly, we confirm that body features can be utilized for affect recognition both in individual settings and in group settings. Therefore, body features can be used for affect recognition while face information is not available due to occlusions etc, which often happens in group settings. Finally, we find that the contextual information of being alone or in a group can be successfully recognized using facial and body cues.

The current works presented in this chapter only utilize hand-crafted features. Methods using hand-crafted features can be easily implemented but they can restrict the representation capability of facial or body expressions to serve different applications in affect analysis and may only capture insignificant characteristics for a task when using hand-crafted features. In contrast, there is no need to define the feature representation format a-priori when using deep architecture methods but instead to learn features from raw image/video data. Deep learning based methods have achieved the-state-of-the-art performance in various challenging computer vision problems, with no exception in affective computing tasks (Li and Deng, 2018). However, to the best of the author's knowledge, there has been no works focusing on investigating affect analysis in group settings using end-to-end deep learning based methods. It would be interesting to have the deep leaning frameworks to learn the affective features directly from the raw images or videos to arrive at the affective states prediction in group settings in an end-to-end manner. On the other hand, though deep learning shows tremendous promise for the object recognition tasks, the deep learning methods can be restricted due to the limited amount of data, which is common for the datasets for affect analysis in group settings, whereas transfer learning and data augmentation methodologies can be used to address this problem.

In this chapter, we present a framework that utilises different face and body features for predicting affect and contextual information, however, we only use vision-based signals, not any physiological signals that are also provided in the dataset. It was shown that the combination of physiological signals and facial expressions can improve the recognition results for the generation of affective tags along valence-arousal space compared to single modality (Koelstra and Patras, 2013). It would be interesting to combine both visual and physiological information for affect analysis in group settings and investigate how different signals influence the affect recognition results, whereas this topic is out of the scope of this thesis as we only focus on visual based affect analysis.

Chapter 5

AFFECT ANALYSIS ACROSS SUBJECTS

5.1 Introduction

In a group setting, there are always multiple people in a scene and individuals in the scene interact with each other in different ways, e.g., communicate verbally, demonstrate facial expressions and make physical contact. Due to the complex dynamics of these interactions, it is very often that one certain person may be occluded by others, which makes direct analysis of his or her affective states (like what we did in Chapter 4) hard or impossible. Instead of direct analysis of one's affective states, in these cases, we can infer and analyse his or her affective states using the information of the other members within the same group. Taking a classroom as one example, when the instructor tells an interesting story, one particular student's facial expression can be inferred as happy from the smiles shown on other students' faces. We term the emotion analysis of one group member using the affective states of other members within the same group as *"cross-subject affect analysis"* or *"affect analysis across subjects"*.

Individuals tend to adapt their behaviours with the other individuals in the same group or in an interaction (Barsade, 2002). The shared information or behaviours among group members provides the possibility of predicting one's affective states using the information of the other group members. In a particular case, we assume that the affect of each subject in a group is synchronised most of the time, i.e., showing similar behaviours. To be more specific, we hypothesise that

- The affect of subjects in the same group are more correlated than that of subjects across different groups.
- It is possible to predict the affect of a subject automatically using the behaviours expressed by the other subject(s) in the same group.

To validate the above hypotheses, in this chapter, we firstly compare the correlation of emotions between subjects from the same group with that between subjects from different groups, and then we conduct experiments to automatically recognise the emotions of one subject using expressive behaviours of the other subject(s) in the same group.

The rest of this chapter is organised as follows: the proposed framework for cross-subject affect analysis in group settings is illustrated in Section 5.2; the experiments and results are presented and discussed in Section 5.3; and conclusions and future work are described in Section 5.4.

5.2 The proposed framework

To investigate how one's affect can be correlated with and predicted from others', we propose frameworks for cross-subject affect analysis along valence and arousal in group videos. An illustration of the motivation of the proposed framework can be seen in Figure 5.1a, which is one of the frames from the GroupDB database detailed in Chapter 4, taken from the AMIGOS dataset (Correa et al., 2018). In this image, we can see that four subjects are watching movies and at this moment all of them are displaying very similar emotion, i.e., positive along valence dimension and high along arousal dimension. In this case, if the information of someone is lost or someone is not willing to share his/her information, the information of the other subject(s) can be utilised to predict the emotion of the lost person, that is the cross-subject affect recognition.

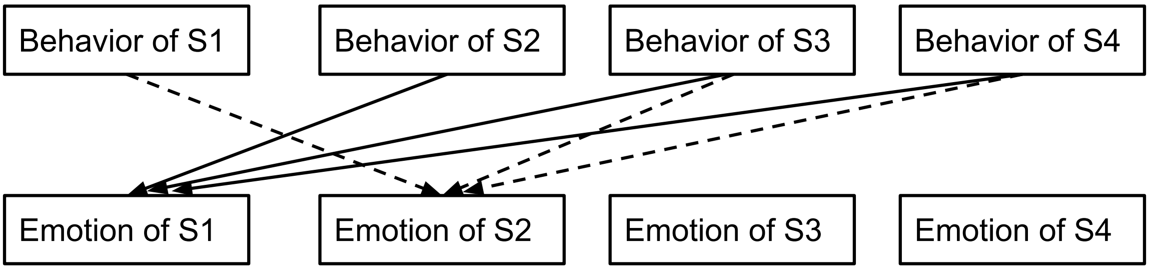
To conceptualise the cross-subject affect analysis, we illustrate the proposed framework in Figure 5.1b. Same as in Figure 5.1a, we take a group of four subjects as an example. In this work, we investigate the cross-subject affect analysis in a pairwise manner, i.e., calculating the correlation of the emotions between two subjects, and predicting the emotion of one individual using the behaviours of only one another. It results into two cases:

- **Cross-subject affect analysis of two subjects in the same group.** For correlation analysis in this case, the correlation between each two subjects within the same group is calculated. For automatic affect recognition in this case, the emotion of the individual is predicted by using another subject in the same group. For example, as shown in Figure 5.1, the facial behaviours expressed by subject 2 (S2), subject 3 (S3) and subject 4 (S4) are separately used to predict the affect of Subject 1 (S1).

- **Cross-subject affect analysis of two subjects in two different groups.** For correlation analysis in this case, one subject is paired by another subject from another group and then the correlation of emotions between these two subjects is calculated. For automatic affect recognition in this case, the emotion of one individual is predicted by using the facial behaviours of another subject from a different group. All of the possible pairs are counted. For example, for subject 1 (S1) from group 1, it is paired with the other 12 subjects from group 2, 3 and 4 separately.

By comparing the correlation and prediction results of these two cases, we can investigate whether the affect of subjects in the same group are more correlated than that of subjects across different groups, whether we can use the information of one subject to predict the

(a) An example of one of the four groups with four subjects watching movies. From this image, we can see that at this moment all of the four subjects are displaying very similar emotion, i.e., positive along valence dimension and high along arousal dimension.



(b) An illustration of the framework for cross-subject affect prediction. For example, while predicting the emotion of Subject 1 (S1) the behaviour of the S2, S3 and S4 is utilised separately; while predicting the emotion of S2, the behaviour of the other three subjects, i.e., S1, S2 or S3, is utilised.

Figure 5.1 An example image from the dataset and an illustration of the proposed framework for cross-subject affect analysis. (a) shows an example image from the group videos and (b) illustrates the framework of the cross-subject affect analysis.

emotion of other subject(s) within the same group, and whether the physical distance between two subjects in the same group has an effect on the affect correlation and affect recognition. The details of the proposed framework are presented as follows.

5.2.1 Correlation analysis of the affect across subjects

To investigate whether the affect of subjects in the same group is more correlated than that of subjects across different groups, we start our analysis from the ground truth level, i.e., the emotion level. There are two cases for analysing the correlation of the affects of two subjects:

- The correlation of subject s and subject m in the group i is calculated and denoted as $C_{i_s i_m}$. Here i denotes the group ID that the subject is from, and $i = 1,2,3,4$, while s and m denote two subjects from group i.

- The correlation of subjects across different groups is denoted as $C_{i_s j_m}$, where i, j denote the group IDs that the subjects are from and $i, j = 1,2,3,4$, and $i \neq j$. s denotes the subject ID of the subject from group i, while m denotes the subject ID of the subject from group j.

We utilise Pearson's Correlation Coefficient (PCC) and Concordance Correlation Coefficient (CCC) (Ringeval et al., 2015) as the evaluation methods.

5.2.2 Automatic affect prediction across subjects in group settings

As in the correlation analysis, the automatic affect prediction across subjects is also divided into two parts:

- To predict the affect of one subject using the facial behaviours expressed by another subject in the same group. The predicted affect in this case is referred as $f_{i_s i_m}$, where i denotes the group of the subject from, and $i = 1,2,3,4$. s and m denote the subject ID.

- To predict the affect of one subject using the facial behaviours of another subject from a different group. The predicted affect in this case is referred as $f_{i_s j_m}$, where i, j denote the group of the subject from and $i, j = 1, 2, 3, 4$, $i \neq j$, and s and m denote the subject ID.

Following the features and temporal models used in Chapter 4, we utilise Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and facial QLZM feature for cross-subject affect regression as shown in Figure 5.2. In the experimental results obtained in Chapter 4, we can see that facial appearance features perform the best for affect analysis among all face and body features that were used, therefore, we utilize facial appearance features, i.e., QLZM, for affect analysis across subjects. We also see in Chapter 4 that the temporal modelling method LSTM outperforms the static learning method SVM. To this end, LSTM plus facial QLZM features are utilised in this chapter. QLZM is extracted from each frame and LSTM is trained using these frame-level facial features, which takes each 20-second clip as a sequence as shown in Figure 5.2. From this figure, we can clearly see that in the cross-subject affect recognition the input sequence is from a subject, but the output is to predict the affect of another subject utilising the information of the input subject. This framework is implemented on PyTorch platform (Paszke et al., 2017).

5.3 Experiments

5.3.1 Experimental data

GroupDB as illustrated in Chapter 4 is used in the experiments in this chapter. Specifically, group videos from four groups are used in the experiments, i.e., three groups (twelve subjects) with recordings of people watching four movies (N1, P1, B1 and U1) and one group (four subjects) with recordings of people watching three movies (B1, N1 and U1). In this case, we have data from sixteen subjects and fifteen sessions in total used in the experiments. One

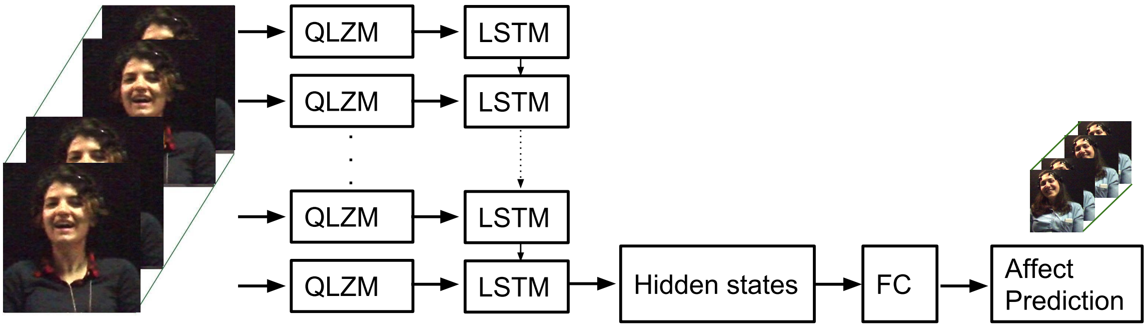


Figure 5.2 Illustration of the approach for cross-subject affect recognition using QLZM with LSTM. There are input sequences from a subject, which are always 20-second clips. Then frame-level QLZM features are extracted. After that, QLZM features extracted from every frame are fed into a one-layer LSTM with 128 hidden states. Finally, affect prediction results can be obtained for the paired subject using the information of the displayed subject in the sequence.

session refers to the recording of one group watching one movie. For each session, 20-second clips in line with the annotations labelled are utilised. The number of the 20-second clips from different sessions varies with the length of the movies, i.e., 70 clips for N1, 70 clips for B1, 56 clips for P1 and 42 clips for U1. As a result, the total number of clips we use in our experiments is $(70(B1) \times 4(4subjects) \times 4(4movies)) + (70(N1) \times 4(4subjects) \times 4(4groups)) + (56(P1) \times 4(4subjects) \times 3(3groups)) + (42(U1) \times 4(4subjects) \times 4(4groups)) = 3,584$.

5.3.2 Experimental results and analysis

Correlation analysis of the affect across subjects

We first conduct the correction analysis of emotions across subjects. The correlation between the emotion levels of two subjects from one group and across different groups are calculated. These emotion levels are the values annotated by the labelers, i.e., the ground truth levels.

These correlation results for the average of all pairs of $C_{i_s i_m}$ and the average of all pairs of $C_{i_s j_m}$ along arousal and valence dimensions are represented in Table 5.1. While calculating the correlations along each dimension between a subject s from group i and a subject m from group j, we use the corresponding ground truth affect when the subject watching the same video segment (i.e., same video, same time). Take the $C_{i_s j_m}$ in terms of PCC along arousal dimension as an example, it can be calculated as follows:

$$C_{i_s j_m} = \frac{cov(X_{i_s}, X_{j_m})}{\sigma_{X_{i_s}} \sigma_{X_{j_m}}}, \tag{5.1}$$

$$cov(X_{i_s}, X_{j_m}) = \frac{1}{N} \sum_{v=1}^{4} \sum_{t=1}^{t_v} ((X_{i_s})_{vt} - \mu_{X_{i_s}})((X_{j_m})_{vt} - \mu_{X_{j_m}}), \tag{5.2}$$

where X_{i_s} denotes the arousal level of a subject s from group i, while X_{j_m} denotes the arousal level of a subject m from group j, where $i \neq j$ and X_{j_m}. v refers to the videos/movies and $v = 1,2,3,4$, corresponding to the four movies people watched. t refers to the 20-second segments along time and the number of segments from video v is denoted as t_v. N is the number of all segments for each subject and $N = \sum_{v=1}^{4} t_v$.

As we mentioned before, the correlation of subjects across different groups is calculated for all pairs: for each subject, it needs to pair with all of the subjects from the other three groups (each time it is paired with one of the subjects from the other three groups). Therefore, in total for one subject, it will get twelve different pairs across three different groups as there are four people in each group. The averages and standard deviations of the correlation results of all these pairs for all subjects are reported in Table 5.1.

From Table 5.1, we can see that the affect of subjects in the same group is much more correlated than that of subjects across different groups. In more details, the affect correlation of subjects in the same group has an average PCC of 0.516 and 0.545 along arousal and valence dimensions respectively, while the affect correlation of subjects across different groups is 0.388 for arousal and 0.390 for valence in terms of PCC. The affect correlation

Table 5.1 The average and the standard deviations of PCC and CCC of all paired subjects in the same group, i.e., all $C_{i_s i_m}$; and the average and the standard deviations of PCC and CCC of all paired subjects across different groups, i.e., all $C_{i_s j_m}$. And the significance tests (p-values) are also reported between $C_{i_s i_m}$ and $C_{i_s j_m}$.

Dimensions	Arousal		Valence	
	PCC	CCC	PCC	CCC
$C_{i_s i_m}$	0.516(0.115)	0.468 (0.130)	0.545 (0.159)	0.498 (0.161)
$C_{i_s j_m}$	0.388(0.154)	0.336(0.152)	0.390(0.203)	0.353(0.194)
$C_{i_s i_m} > C_{i_s j_m}$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$

of subjects in the same group has an average of 0.468 and 0.498 in terms of CCC along arousal and valence respectively, compared to 0.336 and 0.353 for the affect correlation of subjects across different groups. The statistically significant test show that the correlation results of subjects in the same group are significantly higher than those of the subjects across different groups in terms of PCC and CCC along both arousal and valence dimensions, with $p\ value < 0.01$. Taking this finding, we can move forwards to predict the emotion of one subject using the facial behaviours of other members in the same group.

In a further step, we investigate whether the physical distance between two subjects has an effect on the correlation of the affect, where the affect refers to the annotated arousal and valence levels. We present the correlation results of the affect separately for the paired subjects with different physical distances. AMIGOS is in an audience setting, where participants always sit together facing the screen to watch movies. The distance between two subjects in a group g, $S_{g,i}$ and $S_{g,j}$, is defined as $n = |i - j|$ and $n \in \{1,2,3\}$, where $g \in \{1,2,3,4\}$ refers to 4 different groups, and $i, j \in \{1,2,3,4\}$ refer to the IDs of each subject in a group from left to right as shown in Figure 5.3. For example, the distance between $S_{1,1}$ and $S_{1,2}$ is 1, while the distance between $S_{1,1}$ and $S_{1,3}$ is 2. The affect correlation between two subjects in a group g, $S_{g,i}$ and $S_{g,j}$, is defined as $Correlation(AS_{g,i}, AS_{g,j})$. The average of the affect

Group 1	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$
Group 2	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$
Group 3	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$
Group 4	$S_{4,1}$	$S_{4,2}$	$S_{4,3}$	$S_{4,4}$

Figure 5.3 All subjects in four groups.

correlation $Corr_n$ of two subjects with the same distance n is defined as:

$$Corr_n = \frac{1}{Z_n} \sum_{g=1}^{4} \sum_{i=1}^{4} \sum_{j=1}^{4} Correlation(AS_{g,i}, AS_{g,j}) \\ \text{subject to } |i-j| = n \tag{5.3}$$

where Z_n is a normalisation factor equal to the number of pairs with distance n and is defined as:

$$Z_n = \sum_{g=1}^{4} \sum_{i=1}^{4} \sum_{j=1}^{4} 1, \quad \text{subject to } |i-j| = n \tag{5.4}$$

The results obtained for $Corr_1$, $Corr_2$ and $Corr_3$ in terms of PCC and CCC along both arousal and valence dimensions are shown in Table 5.2. From Table 5.2, we can see that the closer the two subjects are staying, the more correlated the affect in terms of arousal and valence is. For example, for arousal dimension the correlation in terms of CCC is 0.500 for $Corr_1$, 0.460 for $Corr_2$, and 0.391 for $Corr_3$; for valence dimension the correlation in terms of CCC is 0.531 for $Corr_1$, 0.501 for $Corr_2$, and 0.391 for $Corr_3$.

Based on these results obtained in Table 5.2, we can see that the physical distance between two subjects has an effect on the correlation of their affect along arousal and valence dimensions. We discuss the results from the following perspectives: seating arrangement, relationship status and the generalisability of the obtained results to different group settings.

Seating arrangement. From the GroupDB, we see that participants sometimes talk to their neighbours making remarks about a movie, and sometimes they look at their neighbours while watching the movie, which communicates their feelings and may explain why two participants seated closer show higher correlations in terms of affect.

Relationship status. During the experiments for the data collection of GroupDB, four participants in each group were self-organised to get seated, therefore, it is possible that participants who know each other well chose to sit closer. There is not enough information to validate the acquaintance between all subjects using GroupDB, but this potentially contributes to the results, i.e., the closer the two participants are, the higher the affect correlation is.

Generalisability to different group settings. The results of the proximity influences we obtained with GroupDB may not be generalisable to other group settings and/or other tasks. For example, (Gedik et al., 2018) did not report that the proximity between people was a factor that triggers synchronous motion for the audience in a live dance performance. Such difference in findings could also be due to our study being an in-lab study with a much smaller group. Their study took place in more “naturalistic” settings and the audience, as well as the room, were much larger. Another practical reason for the difference in findings could be due to the “offline” versus “online” aspect of the stimulus. In our study the movies are recordings, while in theirs the performance is “live”. In the literature, it is reported the audience react differently when they are in a live performance as compared to watching a recorded one. For example, live music engages listeners to a greater extent than pre-recorded music (Swarbrick et al., 2019), and the audience responses to live theatre performance are more cognitive and communicatively effective than the recorded theatre performance (Shrader, 2015). Therefore

Table 5.2 The average and standard deviations (in parentheses) of the affect correlation for two subjects with the same distances $n = \{1, 2, 3\}$ in terms of PCC and CCC are shown here.

Dimensions	Arousal		Valence	
	PCC	CCC	PCC	CCC
$Corr_1$	0.537(0.109)	0.500(0.125)	0.578(0.136)	0.531(0.137)
$Corr_2$	0.512(0.125)	0.460(0.130)	0.543(0.192)	0.501(0.188)
$Corr_3$	0.462(0.107)	0.391(0.127)	0.452(0.133)	0.391(0.143)

audience behaviour in live performances might potentially be less externalised as compared to watching a recorded one.

Automatic affect prediction across subjects

For the automatic affect recognition, it is also divided into two cases, one is to predict the affect of one subject using the behaviours of another subject in the same group and one is to predict the affect of one subject using the behaviours of one subject from a different group, which is in a similar manner to the correlation analysis.

For the cross-subject affect recognition in the same group, the affect of a subject s in group i is predicted using the facial behaviours expressed by another subject m in group i, $f_{i_s i_m}$. Each subject has been paired with all of the other subjects in the same group. For that across different groups, the affect of a subject s in group i is predicted using the facial behaviours expressed by another subject m in group j, $f_{i_s j_m}$. Each subject has been paired with all the subjects in the other three groups. The average and standard deviations of the recognition results of all $f_{i_s i_m}$ and $f_{i_s j_m}$ are represented in Table 5.3 in terms of PCC and CCC along both arousal and valence dimensions. As we can clearly see from Table 5.3, the average results of cross-subject affect recognition for subjects in the same group, i.e., $f_{i_s i_m}$, are much better than that of subjects in different groups, i.e., $f_{i_s j_m}$, in terms of bother PCC and CCC

Table 5.3 The recognition results in terms of PCC and CCC along both arousal and valence dimensions for subjects in the same group (the average and standard deviations of all pairs of $f_{i_s i_m}$) and all paired subjects across different groups (the average and standard deviations of all $f_{i_s j_m}$). And the significance tests (p-values) are also reported between $f_{i_s i_m}$ and $f_{i_s j_m}$.

Dimensions	Arousal		Valence	
	PCC	CCC	PCC	CCC
$f_{i_s i_m}$	0.454(0.133)	0.362(0.140)	0.509(0.146)	0.410(0.150)
$f_{i_s j_m}$	0.259(0.136)	0.194(0.115)	0.243(0.143)	0.190(0.128)
$f_{i_s i_m} > f_{i_s j_m}$	$p < 0.01$	$p < 0.01$	$p < 0.01$	$p < 0.01$

along arousal and valence dimensions. For example, the average PCC between the predicted affect and the ground truth affect obtained with $f_{i_s i_m}$ are 0.454 and 0.509 along arousal and valence dimensions respectively, while that for $f_{i_s j_m}$ is 0.259 and 0.243 corresponding to arousal and valence dimensions. The average CCC obtained with $f_{i_s i_m}$ is 0.362 along arousal dimension and 0.410 along valence dimension, while that for $f_{i_s j_m}$ is 0.194 along arousal dimension and 0.190 along valence dimension. The statistically significant test shows that all the affect recognition results obtained with the subjects in the same group are significantly better than those of the paired subjects across different groups.

The possible reason is that people in the same group share some facial behaviours, that contributes to the affect prediction as we hypothesised. As a result, while the expressive behaviours of a subject is not available due to occlusion or head poses which is one of the main challenges for affect analysis in group settings, the behaviours expressed by the other subject(s) can be used for the affect prediction of that subject.

5.4 Discussion and conclusion

In this chapter, we propose a framework to investigate the cross-subject affect analysis in group videos. We conduct a set of experiments using the GroupDB in AMIGOS database that aims to study affect analysis with a group of people watching stimuli movies. The experimental results show that (1) the affect of subjects in the same group is more correlated than that of subjects across different groups; (2) the affect of a subject predicted using facial behaviours expressed by the other subject in the same group is significantly better than that predicted using the behaviours of a subject in a different group; and (3) the distances between two subjects in the same group have an effect on the emotion correlation. With the above findings, we further validate that people in the same group share some information and are influenced by each other in terms of facial behaviours and emotions. As a result, it is potential to help address one of the main challenges for affect analysis in group settings, i.e., inability to predict facial affect due to occlusion among subjects or due to head pose variations. For example, when the information of one subject is unavailable, we can predict the affect of that subject based on the expressive behaviours of the other subject(s).

In this chapter, we mainly focus on the analysis in a pair-wised manner, i.e., only investigating the influences between two subjects either in the same group or in different groups. It would be interesting to investigate the influences of emotions among more than two people. It would also be interesting to investigate whether it is possible to predict the emotion of one subject using the information of all the other subjects within the same group. How will the results of cross-subject affect analysis with all subjects in one group considered be compared to those obtained by using only one subject? In this manner, we can investigate whether more people will provide complementary information or only redundant information, which will help the analysis of not only the affect but also other social dimensions in group settings.

Chapter 6

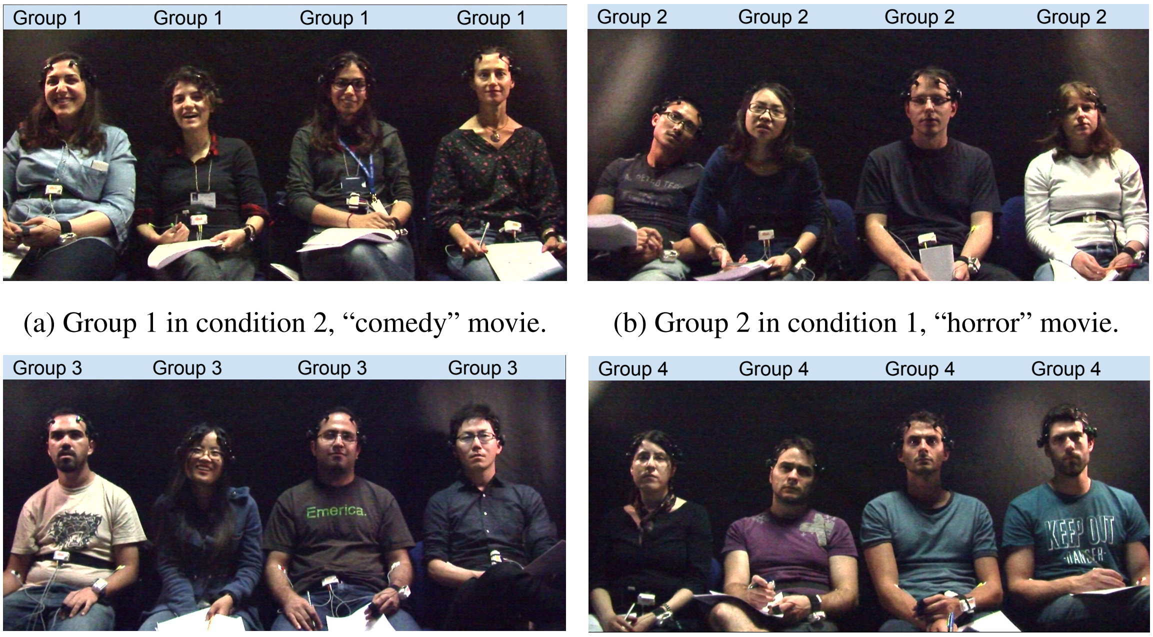
GROUP MEMBERSHIP RECOGNITION IN GROUP VIDEOS

6.1 Introduction

In Chapter 4 and Chapter 5, we have been focusing on investigating the affect recognition in group videos. In addition to affect analysis, nowadays automatic analysis of a group of people from other perspectives has also received much attention in computer vision community for different research purposes. Gallagher *et al.* (Gallagher and Chen, 2009) propose a framework to predict ages and genders of individuals in group images; and Ibrahim *et al.* (Ibrahim et al., 2016) focus on group activity recognition. Research works focusing on the analysis of social dimensions, such as engagement and rapport in group settings have also been reported in (Leite et al., 2015) and (Hagad et al., 2011). Hung *et al.* (Vascon et al., 2016; Zhang and Hung, 2016) proposed methodologies for the detection of free standing conversational groups (also known as F-formation) and for the analysis of social involvement in free standing conversational groups. Most of the aforementioned works analyze what is happening within the group. Recently, works on automatic analysis of the relationship between the members of different groups have emerged and one example is (Mioranda-Correa and Patras, 2018), where Mioranda-Correa *et al.* predict social context, i.e., whether a person is being alone or in a group utilising neuro-physiological signals.

In this chapter we investigate the prediction of group membership for each individual, using vision-based non-verbal behaviours, when they are part of a group of four participants sitting together and watching four movies, i.e., GroupDB presented in Subsection 2.4.3. We form four groups, each of which contains four participants, with no overlaps between the group members (sixteen participants in total). The sixteen participants and the labels of their membership are shown in Figure 6.1. Even though they are performing the same task, individuals in different groups may behave very distinctly due to differences in group composition and dynamics. According to research in cognitive and behavioural science (Barsade, 2002), individuals in a particular group tend to affect the behaviours of each other, i.e., mimic one another or exhibit similarities in non-verbal behaviours. In addition, based on our research in Chapter 4 and Chapter 5, we also found that people in a group influence the behaviours of each other and share some common information. Such shared behaviours within the group, and possible differences among different groups, allow the automatic recognition of group membership (Mou et al., 2016).

Towards this direction, we propose a novel approach to the group membership recognition by introducing a novel *specific recognition model* that is built on the top of a *generic recognition model*. The *generic recognition model* refers to the model that is trained using all data across all different conditions. These conditions can be subjects watching different types of movies, e.g., “horror”, “comedy”, “action”, or “adventure” movies as stated in Table 2.2 in Subsection 2.4.3 and illustrated in 6.1, where different groups are under different conditions. The performance of *generic recognition model* may be significantly limited due to the fact that group members may behave distinctly in different conditions. For example, individuals exhibit very distinct behaviours while watching horror movies compared to the case while watching comedies. To address this issue, one option is to use an *independent recognition*



(a) Group 1 in condition 2, “comedy” movie.

(b) Group 2 in condition 1, “horror” movie.

(c) Group 3 in condition 4, “adventure” movie.

(d) Group 4 in condition 3, “action” movie..

Figure 6.1 Example images for four different groups under different conditions, i.e., watching different movies. The group membership of each subject is corresponding to the group the subject belongs to. For example, the four subjects from (a) group 1, are all labelled as group 1. (a) Group 1, (b) Group 2, (c) Group 3, (d) Group 4. The four groups from (a) to (d) are under conditions, watching “comedy”, “horror”, “adventure”, and“action” movies respectively.

model, i.e., using solely the data from the same condition, but the performance of *independent recognition model* may be restricted due to the limited number of samples available from each condition. To avoid these issues with the *generic recognition model* and the *independent recognition model*, the *specific recognition model* built upon the *generic recognition model* is proposed. The *specific recognition model* is specific to a certain condition, i.e., a certain type of movie, “horror”, “comedy”, “action”, or “adventure”. When the group members are in different conditions, they may react differently; however, they are still part of the same setting performing the same task, i.e., sitting in front of the screen watching movies, which allows them to share some common behavioural characteristics. In light of these, we hypothesise that the *generic recognition model* can provide a useful baseline for the optimisation of the *specific recognition model*. To this end, we propose the *specific recognition model* for each condition specifically, but also we learn it on the top of the *generic recognition model*.

The remaining part of this chapter is organised as follows. The proposed framework for group membership recognition is introduced in Section 6.2; the experiments and results are provided and discussed in Section 6.3; and conclusions and future work are presented in Section 6.4.

6.2 The proposed framework

To solve the problem of group membership recognition, we first propose a two-phase learning framework, where we first train a *generic recognition model* using all videos across all conditions and, then optimise the *specific recognition model* for each specific condition based on the optimisation results obtained from the *generic recognition model*. In the rest part of this thesis, we refer to this two-phase *Specific Recognition Model* as **two-phase SRM**. At a further step, we unify the *generic recognition model* and the *specific recognition model* under a single joint framework. Specifically, we optimise the *generic recognition model* and the *specific recognition model* jointly. In this way, the framework is converted to an end-to-end structure, which is easier for both training and testing. In the rest of the thesis, we refer to the unified model, *Joint Specific Recognition Model*, as the **JointSRM**.

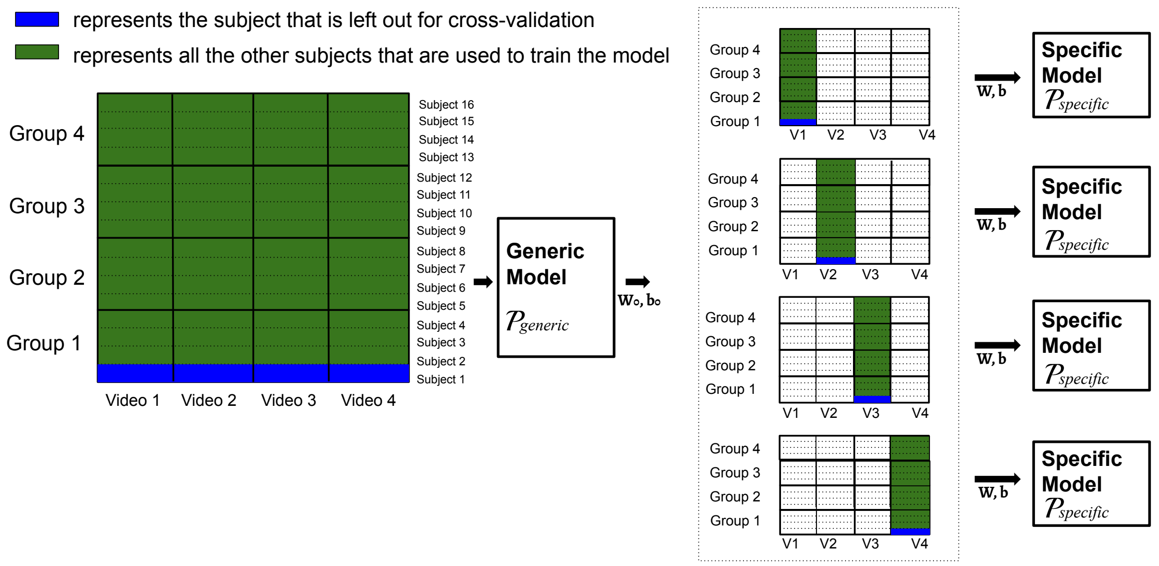
6.2.1 The two-phase learning framework

The proposed two-phase learning framework is illustrated in Figure 6.2, which aims to learn a *specific recognition model* upon a *generic recognition model*. The first step is to learn a *generic recognition model* using all data across all conditions (videos). The second step is to learn the *specific recognition model* using data from only one specific condition based on the optimised *generic recognition model*. As the data across different conditions are all under the same scenario, that is sitting in front of the screen watching movies, we hypothesise that, the two recognition models share some common knowledge and therefore the *generic recognition model* can provide a baseline for optimising the *specific recognition model*.

The generic recognition model

Our recognition models are based on linear Support Vector Machine (SVM). For training each of the linear models we use a Stochastic Gradient Descent (SGD) algorithm (Shalev-Shwartz et al., 2007).

The first step of the proposed framework is to learn the *generic recognition model* using the standard linear SVM. The *generic recognition model* is not taking the different conditions into consideration, but uses all of the available training samples across all conditions. A condition mentioned here refers to a certain movie/video. In this model, we use all of the available training samples, which are from all subjects across all conditions. We denote this training set as $\mathscr{X} = \{(\mathbf{x}_i, z_i), i = 1, \ldots, \ell\}$, where $\mathbf{x}_i$ denotes the i-th training sample and z_i is the corresponding ground truth label, being equal to $+1$ if the sample belongs to the respective positive class, or -1 otherwise.



(a) Generic recognition model

(b) Specific recognition model

Figure 6.2 An illustration of the proposed two-phase SRM. It is divided into two learning phases, i.e., (a) learning the *generic recognition model* and (b) learning the *specific recognition model*. As we apply *leave-one-subject-out* cross-validation, for *generic recognition model*, we leave all of the samples of one subject (blue) out and train the model with all the other samples (green). For the *specific recognition model*, as we have four different videos, we have $n = 4$ specific problems and optimise them based on the optimised weights obtained from the *generic recognition model*. For the specific model, we also do *leave-one-subject-out* cross-validation.

The generic optimisation problem, which we denote as $\mathscr{P}_{generic}$, can be cast as follows:

$$\mathscr{P}_{generic}: \quad \min_{\mathbf{w}_0, b_0} \frac{\lambda}{2}\|\mathbf{w}_0\|^2 + \frac{1}{\ell}\sum_{i=1}^{\ell} \mathscr{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i)), \tag{6.1}$$

where λ is the regularisation parameter and $\mathbf{w}_0$, b_0 are the optimisation parameters. $\mathscr{L}$ denotes the loss function and is given by the hinge-loss, as follows

$$\mathscr{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i)) = \max(0, 1 - z_i(\mathbf{w}_0^\top \mathbf{x}_i + b_0)). \tag{6.2}$$

We use the Pegasos (Shalev-Shwartz et al., 2007) SGD algorithm for solving the above optimisation problem and we arrive at the optimal solution $(\mathbf{w}_0, b_0)$, which describes the optimal hyper-plane $\mathscr{H}_0$: ${\mathbf{w}_0}^\top \mathbf{x} + b_0 = 0$. Then, we use the optimal $\mathbf{w}_0$ to construct the *specific recognition model*, as described below.

The specific recognition model

A *specific recognition model* is specific to a certain condition, i.e., “horror”, “comedy”, “action”, or “adventure”, which is denoted by j in Equation 6.3. The *specific recognition model* is learned utilising the optimisation results obtained from the *generic recognition model*. That is, we use the optimal value for $\mathbf{w}_0$ (by solving the optimisation problem in Equation 6.1) in order to construct the specific optimisation problem.The j-th condition is denoted as $\mathscr{P}^j_{\text{specific}}$ and cast as follows

$$\begin{aligned}\mathscr{P}^j_{\text{specific}}: \; \min_{\mathbf{w}_j, b_j} & \frac{\mu_j}{2}\|\mathbf{w}_j\|^2 + \frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2 \\ & + \frac{1}{\ell_t} \sum_{(\mathbf{x}_i, z_i) \in \mathscr{X}_t} \mathscr{L}(\mathbf{w}_j, b_j; (\mathbf{x}_i, z_i)), \; j = 1, \ldots, 4\end{aligned} \tag{6.3}$$

where $\mathscr{X}_t$ is a subset of the original training set, μ_j and ν_j are regularization parameters, and $\mathscr{L}$ denotes the hinge-loss. The term $\frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2$ is used to bias $\mathbf{w}_j$ to be close to $\mathbf{w}_0$. When

ν is equal to 0, the model becomes the standard linear SVM, while when ν tends to infinity, $\mathbf{w}$ tends to be equal to $\mathbf{w}_0$. The optimal values for μ_j, ν_j are obtained using cross-validation.

For solving $\mathscr{P}^j_{\text{specific}}$, we use a variant of the Pegasos SGD algorithm. The proposed algorithm receives two parameters as input: (1) the number of iterations T, and (2) the number of examples to be used for calculating sub-gradients, k. Initially, we set $\mathbf{w}_j^{(1)}$ to any vector whose norm is at most $\frac{1}{\sqrt{\nu_j}}$ and $b_j^{(1)} = 0$. On the t-th iteration, we randomly choose a subset of $\mathscr{X}$, of cardinality k, i.e., $\mathscr{X}_t \subseteq \mathscr{X}$, where $|\mathscr{X}_t| = k$ and set the learning rate to $\eta_t = \frac{1}{\nu_j t}$. We approximate the objective function of $\mathscr{P}^j_{\text{specific}}$ with

$$\begin{aligned}\mathscr{P}^j_{\text{specific}}: \ \mathscr{J}(\mathbf{w}_j, b_j) = {} & \frac{\mu_j}{2}\|\mathbf{w}_j\|^2 + \frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2 \\ & + \frac{1}{k}\sum_{(\mathbf{x}_i, z_i)\in X_t} \mathscr{L}(\mathbf{w}_j, b_j; (\mathbf{x}_i, z_i)), \ j = 1, \ldots, 4.\end{aligned} \tag{6.4}$$

The update rules are given as follows

$$\mathbf{w}_j^{(t+1)} \leftarrow \mathbf{w}_j^{(t)} - \frac{\eta_t}{k}\frac{\partial \mathscr{J}}{\partial \mathbf{w}_j}, \quad b_j^{(t+1)} \leftarrow b_j^{(t)} - \frac{\eta_t}{k}\frac{\partial \mathscr{J}}{\partial b_j},$$

where the first derivatives of $\mathscr{J}$ with respect to $\mathbf{w}_j$ and b_j are given respectively as

$$\frac{\partial \mathscr{J}}{\partial \mathbf{w}_j} = \mu_j \mathbf{w}_j + \nu_j(\mathbf{w}_j - \mathbf{w}_0) + \frac{1}{k}\sum_{(\mathbf{x}_i, z_i)\in X_t} \frac{\partial \mathscr{L}}{\partial \mathbf{w}_j} \tag{6.5}$$

and

$$\frac{\partial \mathscr{J}}{\partial b_j} = \frac{1}{k}\sum_{(\mathbf{x}_i, z_i)\in X_t} \frac{\partial \mathscr{L}}{\partial b_j}. \tag{6.6}$$

The first derivatives of the hinge loss with respect to $\mathbf{w}_j$ and b_j are given respectively as

$$\frac{\partial \mathscr{L}}{\partial \mathbf{w}_j} = \begin{cases} -z_i \mathbf{x}_i & \text{if } 1 > z_i(\mathbf{w}_j^\top \mathbf{x}_i + b_j), \\ 0 & \text{if } 1 < z_i(\mathbf{w}_j^\top \mathbf{x}_i + b_j). \end{cases} \tag{6.7}$$

and

$$\frac{\partial \mathscr{L}}{\partial b_j} = \begin{cases} -z_i & \text{if } 1 > z_i(\mathbf{w}_j^\top \mathbf{x}_i + b_j), \\ 0 & \text{if } 1 < z_i(\mathbf{w}_j^\top \mathbf{x}_i + b_j). \end{cases} \tag{6.8}$$

Finally, we project $\mathbf{w}_j^{(t+1)}$ onto the ball of radius $\frac{1}{\sqrt{\nu_j}}$, i.e., the set $\mathscr{B} = \{\mathbf{w}_j\colon \|\mathbf{w}_j\| \leq \frac{1}{\sqrt{\nu_j}}$. The output of the algorithm is the pair of $(\mathbf{w}_j^{(T+1)}, b_j^{(T+1)})$.

Once the optimal values of the parameters $\mathbf{w}_j$ and b_j are learned, an unseen testing datum, $\mathbf{x}_t$, can be classified to one of the two classes according to the sign of the (signed) distance between $\mathbf{x}_t$ and the separating hyper-plane. That is, the predicted label of $\mathbf{x}_t$ is computed as $y_t = \operatorname{sgn}(d_t)$, where $d_t = \mathbf{w}_j^\top \mathbf{x}_t + b_j$.

6.2.2 The joint framework

The two-phase framework presented in the above section includes two stages and has to be optimized separately. To simplify the problem, we unify the *generic recognition model* and the *specific recognition model* under a single joint framework. Specifically, in this work we optimize the *generic recognition model* and the *specific recognition model* jointly. In this way, the framework is converted to an end-to-end structure, which is easier for both training and testing. The framework is illustrated in Figure 6.3.

We present a novel joint *specific recognition model* built upon a *generic recognition model*. The Joint Specific Recognition Model (JointSRM) is shown in Equation 6.9.

$$\begin{aligned} \mathscr{P}_{\text{joint-specific}}\colon \min_{\substack{\mathbf{w}_0, b_0\\ \mathbf{w}_j, b_j}} \frac{\lambda}{2}\|\mathbf{w}_0\|^2 + \sum_{j=1}^{4}\left(\frac{\mu_j}{2}\|\mathbf{w}_j\|^2 + \frac{\nu_j}{2}\|\mathbf{w}_j - \mathbf{w}_0\|^2\right) \\ + \frac{1}{\ell}\sum_{i=1}^{\ell}\mathscr{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i)) + \frac{1}{\ell_t}\sum_{j=1}^{4}\left(\sum_{(\mathbf{x}_{it}, z_{it})\in\mathscr{X}_t}\mathscr{L}(\mathbf{w}_j, b_j; (\mathbf{x}_{it}, z_{it}))\right), \end{aligned} \tag{6.9}$$

we denote this training set as $\mathscr{X} = \{(\mathbf{x}_i, z_i), i = 1, \ldots, \ell\}$, where $\mathbf{x}_i$ denotes the feature representation of the i-th training sample and z_i, the corresponding ground truth label,

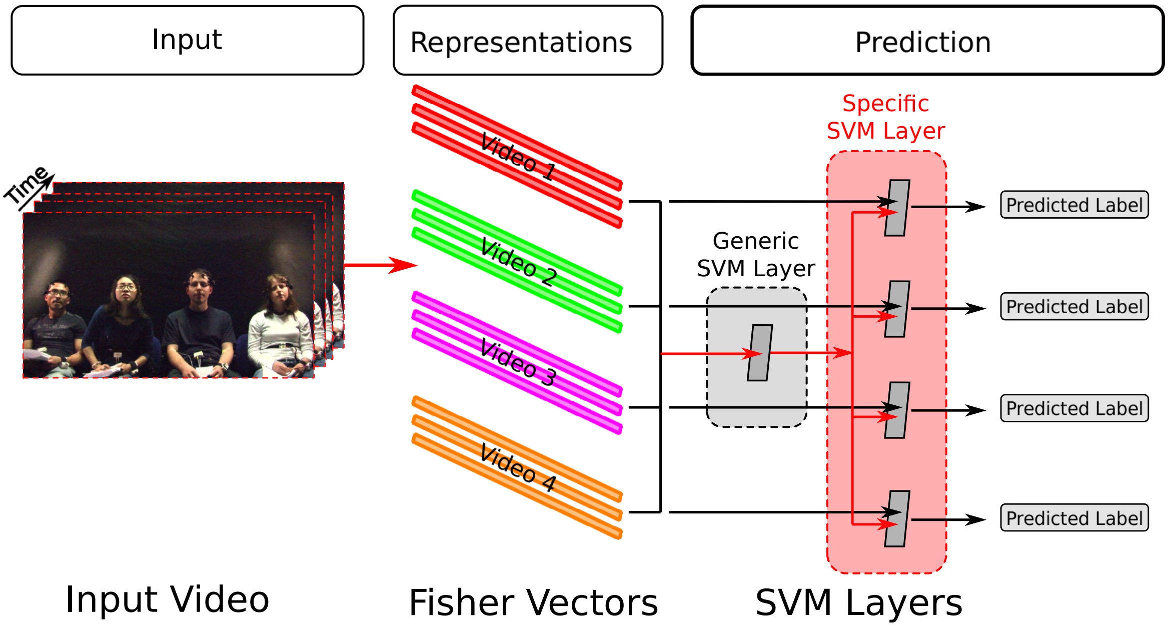


Figure 6.3 An illustration of the proposed framework of JointSRM. It consists of three parts, i.e., input, representations, and prediction. The prediction part contains SVM layers, both generic SVM layer and the specific SVM layers. In this way, we learn the *generic recognition model* in generic SVM layer and learn the *specific recognition model* in specific SVM layer. For the *specific recognition model*, as we have four different conditions, we have $n = 4$ specific problems and optimise them based on the optimised weight, w_0, obtained from the *generic recognition model*. More details of the computation of the loss can refer to Figure 6.4.

being equal to $+1$ if the sample belongs to the respective class, or -1 otherwise. Where $\mathscr{X}_t = \{(\mathbf{x}_{it}, z_{it}), it = 1, \ldots, \ell_t\}$ is a subset of the original training set, $\mathbf{w}_0$, b_0, $\mathbf{w}_j$, b_j are the optimisation parameters (for the generic and the j-th specific model, respectively), λ, μ_j, and ν_j, $j = 1, \ldots, 4$ are regularisation hyper-parameters, and $\mathscr{L}$ denotes the hinge-loss.

For the two-phase SRM presented in Subsection 6.2.1, the *generic recognition model* and the various *specific recognition models* were trained separately. Specifically, we first trained a *generic recognition model*, obtaining an optimal value of the parameter $\mathbf{w}_0$, and then we trained a set of *specific recognition models* based on the optimised *generic recognition model*. For the JointSRM, it is an end-to-end approach to train the *generic recognition model*

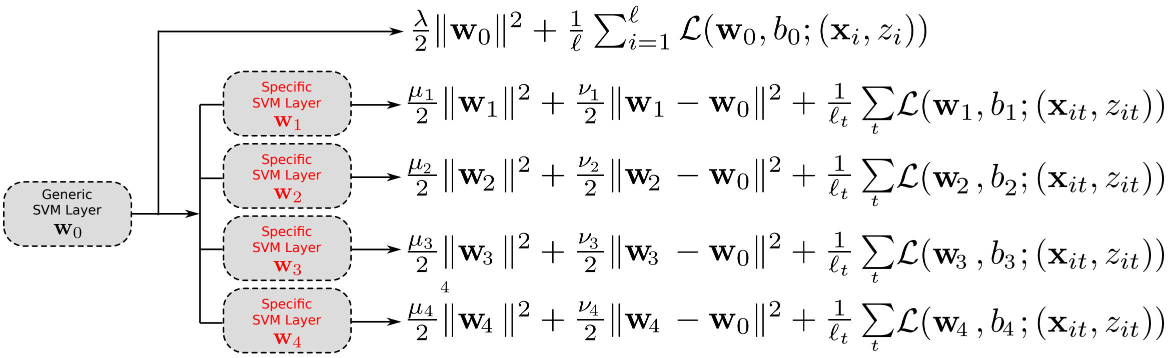

Generic
SVM Layer
$\mathbf{w}_0$
Specific
SVM Layer
$\mathbf{w}_1$
Specific
SVM Layer
$\mathbf{w}_2$
Specific
SVM Layer
$\mathbf{w}_3$
Specific
SVM Layer
$\mathbf{w}_4$
$\frac{\lambda}{2}\|\mathbf{w}_0\|^2 + \frac{1}{\ell}\sum_{i=1}^{\ell}\mathcal{L}(\mathbf{w}_0, b_0; (\mathbf{x}_i, z_i))$
$\frac{\mu_1}{2}\|\mathbf{w}_1\|^2 + \frac{\nu_1}{2}\|\mathbf{w}_1 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t \mathcal{L}(\mathbf{w}_1, b_1; (\mathbf{x}_{it}, z_{it}))$
$\frac{\mu_2}{2}\|\mathbf{w}_2\|^2 + \frac{\nu_2}{2}\|\mathbf{w}_2 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t \mathcal{L}(\mathbf{w}_2, b_2; (\mathbf{x}_{it}, z_{it}))$
$\frac{\mu_3}{2}\|\mathbf{w}_3\|^2 + \frac{\nu_3}{2}\|\mathbf{w}_3 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t \mathcal{L}(\mathbf{w}_3, b_3; (\mathbf{x}_{it}, z_{it}))$
4
$\frac{\mu_4}{2}\|\mathbf{w}_4\|^2 + \frac{\nu_4}{2}\|\mathbf{w}_4 - \mathbf{w}_0\|^2 + \frac{1}{\ell_t}\sum_t \mathcal{L}(\mathbf{w}_4, b_4; (\mathbf{x}_{it}, z_{it}))$

Figure 6.4 Illustration of the computation of the loss for the JointSRM model.

and all the *specific recognition models* simultaneously, simplifying the whole procedure significantly.

6.3 Experiments

Experiments are conducted using a database collected to study group analysis from multimodal cues while each group with four participants were watching a number of long movie segments (Correa et al., 2018), i.e., GroupDB presented in Subsection 2.4.3. They were arranged into four groups with four participants in each group watching all of the four videos listed in Table 2.2 in Subsection 2.4.3 together. This dataset contains data from four groups with recordings under four different conditions, i.e., “horror”, “comedy”, “action”, and “adventure”, as illustrated in Table 2.2 in Subsection 2.4.3. Four frames from four different conditions are shown in Figure 6.1. Group videos from four groups are used in the experiments, i.e., three groups (twelve subjects) with recordings of people watching four movies (N1, P1, B1 and U1) and one group (four subjects) with recordings of people watching three movies (B1, N1 and U1). In this case, we have data from sixteen subjects and fifteen sessions in total used in the experiments. One session refers to the recording of one group watching one movie. For each session, 20-second clips in line with the annotations labelled are utilised. The number of the 20-second clips from different sessions varies with the length of the movies, i.e., 70 clips for N1, 70 clips for B1, 56 clips for P1 and 42 clips for U1. As a result, the total number of clips we use in our experiments is $(70(B1) \times 4(4subjects) \times 4(4groups)) + (70(N1) \times 4(4subjects) \times 4(4groups)) + (56(P1) \times 4(4subjects) \times 3(3groups)) + (42(U1) \times 4(4subjects) \times 4(4gr-oups)) = 3,584$.

6.3.1 Implementation details

The network of JointSRM is implemented using Theano (Theano Development Team, 2016) and Lasagne (Dieleman et al., 2015) libraries. All the parameters of the network, i.e., for the generic SVM layer and the four specific SVM layers as shown in Figure 6.3, are learned using the standard back-propagation technique. In terms of feature representations, we use the body HOF features for group membership recognition as some pilot experiments showed that body HOF features outperform the other facial and body features for group membership recognition.

On one hand, we compare the proposed *specific recognition model* with two other models, (1) the *generic recognition model* that trained across all different conditions and (2) the *independent recognition model* that trained directly in each specific condition. We also compare this JointRSM to the two-phase SRM.

In order to avoid subject-dependency problem, group membership recognition models are trained by applying *leave-one-subject-out* cross-validation. *Leave-one-subject-out* refers to, in each fold, using eleven subjects for training-validation and the remaining one subject for testing. Each time the parameters of the model are optimised over the training-validation samples. The experimental results of the membership recognition are evaluated by the recognition accuracy. In addition, we perform statistical significance analysis to see the significance of the results obtained.

Table 6.1 Group membership recognition results with both two-phase SRM and JointSRM using different models, the proposed *specific recognition model*, *generic recognition model* and *independent recognition model*. The average recognition accuracy of all subjects obtained from *leave-one-subject-out* cross-validation and the standard deviations among all subjects are provided.

Different Models	Acc (std) chance level=25% **SRM**	Acc (std) chance level=25% **JointSRM**
Generic recognition model ($\nu \to \infty$)	26% (18%)	26% (18%)
Independent recognition model ($\nu = 0$)	30% (14%)	30% (14%)
Specific recognition model	**38% (20%)**	**44% (21%)**

6.3.2 Results and analysis

The recognition results in terms of recognition accuracy by applying *leave-one-subject-out* cross-validation are shown in Table 6.1. From Table 6.1, we can clearly see that the proposed *specific recognition model* outperforms the other two models in terms of recognition accuracy under both two-phase SRM and JointSRM setups. A recognition accuracy of 38% is obtained for the *specific recognition model* tested on two-phase SRM, while 26% and 30% are obtained from *generic recognition model* and *independent recognition model* respectively. A recognition accuracy of 44% is obtained for the *specific recognition model* tested on two-phase SRM, while 26% and 30% are obtained from *generic recognition model* and *independent recognition model* respectively. A recognition accuracy of 44% is obtained for the *specific recognition model* with tested on JointSRM, while 26% and 30% are obtained from *generic recognition model* and *independent recognition model* respectively. We also perform a t-test to see the statistical significance, which is also listed in Table 6.1. The statistical significance tests show that the results obtained with the proposed *specific recognition model* are significantly better than chance level, but not for *generic recognition model* and *independent recognition model*.

We also compared the performance obtained with the *specific recognition model* between the two-phase SRM and the JointSRM. we can see that a recognition accuracy of 44% is obtained for the *specific recognition model* with JointSRM, while 38% is obtained for the *specific recognition model* with two-phase SRM. In addition, the joint framework can be trained more easily compared to the non-joint framework, which needs to be trained by two steps, first *generic recognition model* and then *specific recognition model*. However, the joint framework can be trained in one step, which can simplify the problem in terms of implementation but provide better results. The computational cost for training two-phase SRM and JointSRM models in terms of time is 28,570 seconds and 4,050 seconds respectively while implementing on a computer with with 32G RAM and Intel Core i7-4790S CPU. Although the cost is much lower for JointSRM, we have to bear in mind that they are not directly comparable as JointSRM has been trained in a GPU mode with a Titan X GPU used.

6.4 Discussion and conclusion

In this chapter, we propose a novel framework for group membership recognition in group settings. To achieve it, we propose a novel *specific recognition model* that is learned jointly with a *generic recognition model*. To optimise the *specific recognition model*, we propose two different approaches, i.e., the two-phase SRM and the JointSRM. We conduct a set of experiments for group membership recognition on GroupDB that include different groups, with each group comprising four participants watching four different types of movies. The experimental results show that the proposed *specific recognition model* outperforms the compared approaches, i.e., *generic recognition model*. In addition, compared to two-phase SRM, JointSRM can be trained at once by learning both the *generic recognition model* and all the *specific recognition models* simultaneously, rather than learning them separately. In this way, the framework for JointSRM is simplified, while at the same time its performance is improved. Furthermore, as group membership can be recognised using non-verbal behaviours, i.e., body behaviours, it indicates that individuals affect each other's behaviours within a group and their nonverbal behaviors share commonalities. Our results also show that capitalising on shared information in a generic recognition problem is important for learning the specific problem at hand, and this optimisation approach can be possibly transferred to other recognition domains.

Despite the promising results obtained in the experiments, analysis of group membership remains a challenging problem. It would be interesting to experiment with other feature representations for group membership recognition. It is also important to use different contextual information to assist the recognition process, such as personality, movie preference, and the personal relationships between group members. In addition, this learning approach from generic to specific is potential to be applied to other recognition problems, such as affect recognition and engagement recognition.

Chapter 7

CONCLUSIONS

This chapter concludes this thesis with an overview of the presented research and provides guidance towards possible future directions.

7.1 Summary of findings and achievements

In this thesis, we have described our research on affect analysis and group membership recognition in group videos. The detailed contributions are summarized below.

- In Chapter 4, we investigate the affect analysis in individual and group settings. To this end, a framework is proposed using different facial and body behaviours, which shows that the method used for affect analysis in individual settings can be transferred to the affect recognition of individuals in group settings. Among different facial and body features, the proposed vQLZM features have been found to perform the best for predicting affective states of individuals among different unimodal features. In addition, the contextual information of being alone or in-a-group can be successfully predicted using facial and body cues.

- In Chapter 5, a novel framework for affect analysis across subjects in group videos is proposed. It analyses the correlation of the affect among group members and presents affect recognition results of one subject using the behaviours expressed by another subject in the same group. A set of experiments are conducted using group videos, i.e., GroupDB. The experimental results show that (1) the affect of subjects in the same group is more correlated than that of subjects across different groups and (2) the affect of a subject predicted using facial behaviours expressed by the other subject in the same group is significantly better than that predicted using the behaviours of a subject in a different group. The cross-subject emotion recognition is expected to help address one of the main challenges for affect analysis in group settings, i.e., inability to predict facial affect due to occlusion among subjects or due to head pose variations. When the information of one subject is unavailable, we can predict the affect of that subject based on the expressive behaviours of the other subject(s).

- In Chapter 6, a novel framework for group membership recognition is introduced, i.e., the *specific recognition model*, that is built on the top of a *generic recognition model*. For group membership recognition, we use GroupDB that includes four different groups watching different types of movies, i.e., “horror”, “comedy”, “action”, and “adventure”, which are taken as four different conditions. The *generic recognition model* is trained using all data across all conditions, however, since group members may behave distinctly in different conditions (while watching different types of movies), the performance of *generic recognition model* is limited. To address this, the *specific recognition model* is proposed for each specific condition and built on the top of the *generic recognition model*, so as to use the *generic recognition model* to provide a baseline. For optimisation, a two-phase optimisation method is first proposed, where the *specific recognition model* is learned after the *generic recognition model*. And then in order to simplify the optimisation procedure, a JointSRM is proposed to learn the *specific recognition model* and the *generic recognition model* jointly. We conducted a number of experiments and found that the proposed *specific recognition model* outperforms the compared approaches, i.e., *generic recognition model*. In addition, compared to two-phase SRM, JointSRM improves the recognition accuracy of group membership and can be trained more easily learning both the *generic recognition model* and all the *specific recognition models* simultaneously.

7.2 Directions for future works

In the field of affect analysis in group settings, there are still a number of open questions to address.

- The group dataset used in this thesis is limited to the audience scenario with fixed number of people in the scene. It is in need to collect group datasets in more naturalistic scenarios, such as in a group interaction setting, where group members interact with each other, and there will be new people joining the group and some leaving the group. The affect analysis in such settings will be more challenging due to the various dynamics, but provide more research directions and will advance human robot interaction further.

- It is challenging to collect a large dataset for affect analysis due to that it is expensive and time consuming to collect and annotate emotions especially in group settings with multiple people in a scene. Therefore, it will be helpful to use generative models, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Karras et al., 2018) and the flow-based generative models (Kingma and Dhariwal, 2018), to generate images and videos for affect analysis in group settings. Huang *et al.* (Huang and Khan, 2017; 2018) propose frameworks using GANs to generate facial expressions in dyadic scenarios, which may be extended to group settings.

- Multitask learning (Caruana, 1997) can be utilised to learn arousal and valence dimensions jointly. Related tasks often have inter-dependencies on each other. Multitask learning aims to utilise the inter-dependencies among the related tasks to help learn each task better. Currently, all methods used in the thesis are conducting arousal and valence recognition separately. However, arousal and valence are related to each other - when people feel more positive or negative, they tend to show higher arousal in general (Kuppens et al., 2013). In this case, multitask learning that learns the emotional attributes jointly along arousal and valence dimensions should outperform the recognition results obtained for arousal and valence dimensions separately.

- As the expression of emotion is influenced by various factors, such as personality (Keltner, 2003) and cultures (Matsumoto, 1991; 1989), it would be an interesting question to investigate whether such information can be utilised for the recognition of affective states. Especially in group settings, people with different personalities may play different social roles and behave very differently. In the current GroupDB, personality of the participants are also annotated by self-assessment, which may be used as a hidden state to help improve the accuracy of affect recognition.

- In this thesis, we present a multi-modal framework that utilises different face and body features, however, we only use vision-based signals and have never touched the physiological signals that are provided in the dataset. It has been shown that the combination of physiological signals and facial expressions can improve the recognition results for the generation of affective tags along valence-arousal space compared to single modality (Koelstra and Patras, 2013). It would be interesting to combine both visual and physiological information for affect analysis in group settings.

7.3 Closing remarks

As we saw in Plato's quote at the beginning of the thesis, human behaviour flows from three main sources: desire, emotion and knowledge. In this thesis, we have discussed how to analyse one of the sources, i.e., emotion, extensively. However, we have not touched much on the other two sources, namely, desire and knowledge. These three sources interconnect with each other. Desire is a sense of longing or hoping for a person, object, or outcome. When a person desires something or someone, their sense of longing is excited by the enjoyment or the thought of the item or person. Emotions are the key element in decision making and gaining knowledge, and central to the process of rational thought (Spence, 1995). Without emotions to guide one's intelligence, logical decisions cannot be made and knowledge base cannot be built (Spence, 1995). As the three sources are all important and closely related to each other, it is important and interesting to find approaches to investigate them together that the author believes could advance the human-robot interaction in a further step.

Bibliography

Abadi, M. K., Subramanian, R., Kia, S. M., Avesani, P., Patras, I. and Sebe, N. (2015), 'Decaf: Meg-based multimodal database for decoding affective physiological responses', *IEEE Transactions on Affective Computing* **6**(3), 209–222.

Afshar, S. and Ali Salah, A. (2016), Facial expression recognition in the wild using improved dense trajectories and fisher vector encoding, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)'.

Ahonen, T., Hadid, A. and Pietikainen, M. (2006), 'Face description with local binary patterns: Application to face recognition', *IEEE Transactions on Pattern Analysis & Machine Intelligence* (12), 2037–2041.

Aran, O. and Gatica-Perez, D. (2010), Fusing audio-visual nonverbal cues to detect dominant people in group conversations, *in* 'Proceedings of IEEE International Conference on Pattern Recognition (ICPR)'.

Aran, O. and Gatica-Perez, D. (2013), One of a kind: Inferring personality impressions in meetings, *in* 'Proceedings of ACM International Conference on Multimodal Interaction (ICMI)'.

Aran, O., Hung, H. and Gatica-Perez, D. (2010), 'A multimodal corpus for studying dominance in small group conversations', *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality* **22**, 22–26.

Arunnehru, J. and Geetha, M. K. (2017), Automatic human emotion recognition in surveillance video, *in* 'Intelligent Techniques in Signal Processing for Multimedia Security'.

Avci, U. and Aran, O. (2014), Effect of nonverbal behavioral patterns on the performance of small groups, *in* ‘Proceedings of ACM workshop on Understanding and Modeling Multiparty on International Conference on Multimodal Interaction’.

Avci, U. and Aran, O. (2016), ‘Predicting the performance in decision-making tasks: From individual cues to group interaction’, *IEEE Transactions on Multimedia* **18**(4), 643–658.

Bänziger, T., Mortillaro, M. and Scherer, K. R. (2012), ‘Introducing the geneva multimodal expression corpus for experimental research on emotion perception.’, *Emotion* **12**(5), 1161.

Barros, P., Weber, C. and Wermter, S. (2015), Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction, *in* ‘Proceedings of IEEE-RAS International Conference on Humanoid Robots (Humanoids)’.

Barsade, S. G. (2002), ‘The ripple effect: Emotional contagion and its influence on group behavior’, *Administrative Science Quarterly* **47**(4), 644–675.

Barsäde, S. G. and Gibson, D. E. (1998), ‘Group emotion: A view from top and bottom’, *Research on managing groups and teams* .

Barsade, S. G. and Gibson, D. E. (2012), ‘Group affect its influence on individual and group outcomes’, *Current Directions in Psychological Science* **21**(2), 119–123.

Bartlett, M. S., Littlewort, G. C., Frank, M. G. and Lee, K. (2014), ‘Automatic decoding of facial movements reveals deceptive pain expressions’, *Current Biology* **24**(7), 738–743.

Bartlett, M. S., Littlewort, G., Fasel, I. and Movellan, J. R. (2003), Real time face detection and facial expression recognition: Development and applications to human computer interaction., *in* ‘Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)’.

Bon, G. L. (1896), ‘The crowd: A study of the popular mind’.

Cai, H. and Lin, Y. (2011), ‘Modeling of operators’ emotion and task performance in a virtual driving environment’, *International Journal of Human-Computer Studies* **69**(9), 571–586.

Caruana, R. (1997), ‘Multitask learning’, *Machine learning* **28**(1), 41–75.

Celiktutan, O. and Gunes, H. (2014), Continuous prediction of perceived traits and social dimensions in space and time, *in* ‘Proceedings of International Conference on Image Processing (ICIP)’.

Celiktutan, O. and Gunes, H. (2017), ‘Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability’, *IEEE Transactions on Affective Computing* **8**(1), 29–42.

Chang, C.-C. and Lin, C.-J. (2011), ‘LIBSVM: A library for support vector machines’, *ACM Transactions on Intelligent Systems and Technology* **2**(3), 27.

Chen, S., Jin, Q., Zhao, J. and Wang, S. (2017), Multimodal multi-task learning for dimensional and continuous emotion recognition, *in* ‘Proceedings of International Workshop on Audio/Visual Emotion Challenge’.

Chen, S., Tian, Y., Liu, Q. and Metaxas, D. N. (2013), ‘Recognizing expressions from face and body gesture by temporal normalized motion and appearance features’, *Image and Vision Computing* **31**(2), 175–185.

Chung, S., Lount Jr, R. B., Park, H. M. and Park, E. S. (2018), ‘Friends with performance benefits: A meta-analysis on the relationship between friendship and group performance’, *Personality and Social Psychology Bulletin* **44**(1), 63–79.

Cohn, J. F. and De la Torre, F. (2014), ‘Automated face analysis for affective computing’, *In the Oxford handbook of affective computing* pp. 131–150.

Correa, J. A. M., Abadi, M. K., Sebe, N. and Patras, I. (2018), ‘Amigos: A dataset for affect, personality and mood research on individuals and groups’, *IEEE Transactions on Affective Computing* (doi: 10.1109/TAFFC.2018.2884461).

Correa, J. A. M. (2018), ‘Phd thesis: Personality, mood and affect recognition based on neurophysiological signals for individuals and groups’, *Queen Mary University of London* .

Cronbach, L. J. (1951), ‘Coefficient alpha and the internal structure of tests’, *psychometrika* **16**(3), 297–334.

Cruz, A., Bhanu, B. and Yang, S. (2011), A psychologically-inspired match-score fusion model for video-based facial expression recognition, *in* ‘Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)’.

Dalal, N. and Triggs, B. (2005), Histograms of oriented gradients for human detection.

Dautenhahn, K. (2007), ‘Socially intelligent robots: dimensions of human–robot interaction’, *Philosophical Transactions of the Royal Society B: Biological Sciences* **362**(1480), 679–704.

Dhall, A. and Goecke, R. (2015), A temporally piece-wise fisher vector approach for depression analysis, *in* ‘Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)’.

Dhall, A., Goecke, R. and Gedeon, T. (2015), ‘Automatic group happiness intensity analysis’, *IEEE Transactions on Affective Computing* **6**(1), 13–26.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J. and Gedeon, T. (2017), From individual to group-level emotion recognition: Emotiw 5.0, *in* ‘Proceedings of ACM International Conference on Multimodal Interaction (ICMI)’.

Dhall, A., Goecke, R., Joshi, J., Hoey, J. and Gedeon, T. (2016), Emotiw 2016: video and group-level emotion recognition challenges, *in* ‘Proceedings of ACM International Conference on Multimodal Interaction (ICMI)’.

Dhall, A., Joshi, J., Radwan, I. and Goecke, R. (2012), Finding happiest moments in a social context, *in* ‘Proceedings of Asian Conference on Computer Vision (ACCV)’.

Dhall, A., Joshi, J., Sikka, K., Goecke, R. and Sebe, N. (2015), The more the merrier: Analysing the affect of a group of people in images, *in* ‘Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)’.

Dhall, A., Kaur, A., Goecke, R. and Gedeon, T. (2018), Emotiw 2018: Audio-video, student engagement and group-level affect prediction, *in* 'Proceedings of ACM International Conference on Multimodal Interaction (ICMI)'.

Dhall, A. et al. (2012), 'Collecting large, richly annotated facial-expression databases from movies', *IEEE MultiMedia* **19**(3), 34–41.

Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S. K., Nouri, D. et al. (2015), 'Lasagne: First release.'.
URL: *http://dx.doi.org/10.5281/zenodo.27878*

Du, S., Tao, Y. and Martinez, A. M. (2014), 'Compound facial expressions of emotion', *Proceedings of the National Academy of Sciences* .

Ekman, P. and Friesen, W. V. (2003), *Unmasking the face: A guide to recognizing emotions from facial clues*, Ishk.

Fabian Benitez-Quiroz, C., Srinivasan, R. and Martinez, A. M. (2016), Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'.

Forsyth, D. R. (2018), *Group dynamics*, Cengage Learning.

Friesen, E. and Ekman, P. (1978), 'Facial action coding system: a technique for the measurement of facial movement', *Palo Alto* **3**.

Gallagher, A. and Chen, T. (2009), Understanding images of groups of people, *in* 'Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)'.

Gatica-Perez, D. (2009), 'Automatic nonverbal analysis of social interaction in small groups: A review', *Image and vision computing* **27**(12), 1775–1787.

Gedik, E., Cabrera-Quiros, L., Martella, C., Englebienne, G. and Hung, H. (2018), 'Towards analyzing and predicting the experience of live performances with wearable sensing', *IEEE Transactions on Affective Computing* **14**(8).

Ghosh, S., Dhall, A. and Sebe, N. (2018), Automatic group affect analysis in images via visual attribute and feature networks, *in* 'Proceedings of IEEE International Conference on Image Processing (ICIP)'.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H. et al. (2013), Challenges in representation learning: A report on three machine learning contests, *in* 'Proceedings of International Conference on Neural Information Processing'.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014), Generative adversarial nets, *in* 'Proceedings of Conference on Advances in Neural Information Processing Systems (NIPS)'.

Gross, R., Matthews, I., Cohn, J., Kanade, T. and Baker, S. (2010), 'Multi-pie', *Image and Vision Computing* **28**(5), 807–813.

Gunes, H. and Pantic, M. (2010), 'Automatic, dimensional and continuous emotion recognition', *International Journal of Synthetic Emotions* **1**(1), 68–99.

Gunes, H. and Piccardi, M. (2005), Fusing face and body gesture for machine recognition of emotions, *in* 'Proceedings of IEEE International Workshop on Robot and Human Interactive Communication'.

Gunes, H. and Piccardi, M. (2007), 'Bi-modal emotion recognition from expressive face and body gestures', *Journal of Network and Computer Applications* **30**(4), 1334–1345.

Gunes, H. and Piccardi, M. (2008), 'Automatic temporal segment detection and affect recognition from face and body display', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**(1), 64–84.

Gunes, H. and Schuller, B. (2013), ‘Categorical and dimensional affect analysis in continuous input: Current trends and future directions’, *Image and Vision Computing* **19**(3), 34–41.

Gunes, H., Shan, C., Chen, S. and Tian, Y. (2015), ‘Bodily expression for automatic affect recognition’, *Emotion recognition: A pattern analysis approach* pp. 343–377.

Guo, X., Zhu, B., Polanía, L. F., Boncelet, C. and Barner, K. E. (2018), Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions, *in* ‘Proceedings of ACM International Conference on Multimodal Interaction (ICMI)’.

Hagad, J. L., Legaspi, R., Numao, M. and Suarez, M. (2011), Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence, *in* ‘Proceedings of Asian Conference on Computer Vision (ACCV)’.

Happy, S. and Routray, A. (2015), ‘Automatic facial expression recognition using features of salient facial patches’, *IEEE Transactions on Affective Computing* **6**(1), 1–12.

Hernandez, J., Hoque, M., Drevo, W. and Picard, R. W. (2012), ‘Mood meter: counting smiles in the wild’, *Association for Computing Machinery* .

Hess, U., Banse, R. and Kappas, A. (1995), ‘The intensity of facial expression is determined by underlying affective state and social situation.’, *Journal of personality and social psychology* **69**(2), 280.

Hochreiter, S. and Schmidhuber, J. (1997), ‘Long short-term memory’, *Neural computation* **9**(8), 1735–1780.

Horn, B. K. and Schunck, B. G. (1981), ‘Determining optical flow’, *Artificial intelligence* **17**(1-3), 185–203.

Huang, D., Shan, C., Ardabilian, M., Wang, Y. and Chen, L. (2011), ‘Local binary patterns and its application to facial image analysis: a survey’, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **41**(6), 765–781.

Huang, X., Dhall, A., Goecke, R., Pietikainen, M. and Zhao, G. (2018), 'Multi-modal framework for analyzing the affect of a group of people', *IEEE Transactions on Multimedia* **20**(10), 2706–2721.

Huang, X., Dhall, A., Zhao, G., Goecke, R. and Pietikäinen, M. (2015), Riesz-based volume local binary pattern and a novel group expression model for group happiness intensity analysis., *in* 'Proceedings of British Machine and Vision Conference (BMVC)'.

Huang, Y. and Khan, S. (2018), A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions, *in* 'Proceedings of ACM International Conference on Multimodal Interaction (ICMI)'.

Huang, Y. and Khan, S. M. (2017), Dyadgan: Generating facial expressions in dyadic interactions, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)'.

Hung, H. and Gatica-Perez, D. (2010), 'Estimating cohesion in small groups using audio-visual nonverbal behavior', *IEEE Transactions on Multimedia* **12**(6), 563–575.

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A. and Mori, G. (2016), A hierarchical deep temporal model for group activity recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'.

Jain, V., Crowley, J. L., Dey, A. K. and Lux, A. (2014), Depression estimation using audiovisual features and fisher vector encoding, *in* 'Proceedings of International Workshop on Audio/Visual Emotion Challenge'.

Jiang, B., Valstar, M. F., Martinez, B. and Pantic, M. (2014), 'A dynamic appearance descriptor approach to facial actions temporal modeling.', *IEEE Transactions Cybernetics* **44**(2), 161–174.

Jiang, B., Valstar, M. F. and Pantic, M. (2011), Action unit detection using sparse appearance descriptors in space-time video volumes, *in* 'Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)'.

Joshi, J., Dhall, A., Goecke, R., Breakspear, M., Parker, G. et al. (2012), Neural-net classification for spatio-temporal descriptor based depression analysis., *in* ‘Proceedings of IEEE International Conference on Pattern Recognition (ICPR)’.

Junek, W. (2007), ‘Mind reading: The interactive guide to emotions.’.

Jung, H., Lee, S., Yim, J., Park, S. and Kim, J. (2015), Joint fine-tuning in deep neural networks for facial expression recognition, *in* ‘Proceedings of IEEE International Conference on Computer Vision (ICCV)’.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N. et al. (2016), ‘Emonets: Multimodal deep learning approaches for emotion recognition in video’, *Journal on Multimodal User Interfaces* **10**(2), 99–111.

Kaltwang, S., Rudovic, O. and Pantic, M. (2012), Continuous pain intensity estimation from facial expressions, *in* ‘Proceedings of International Symposium on Visual Computing’.

Kanade, T., Tian, Y. and Cohn, J. F. (2000), Comprehensive database for facial expression analysis, *in* ‘Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)’.

Karg, M., Samadani, A.-A., Gorbet, R., Kühnlenz, K., Hoey, J. and Kulić, D. (2013), ‘Body movements for affective expression: A survey of automatic recognition and generation’, *IEEE Transactions on Affective Computing* **4**(4), 341–359.

Karras, T., Laine, S. and Aila, T. (2018), ‘A style-based generator architecture for generative adversarial networks’, *arXiv preprint arXiv:1812.04948* .

Kaya, H., Gürpinar, F., Afshar, S. and Salah, A. A. (2015), Contrasting and combining least squares based learners for emotion recognition in the wild, *in* ‘Proceedings of ACM International Conference on Multimodal Interaction (ICMI)’.

Kelly, J. R. and Barsade, S. G. (2001), ‘Mood and emotions in small groups and work teams’, *Organizational behavior and human decision processes* **86**(1), 99–130.

Keltner, D. (2003), 'Expression and the course of life: Studies of emotion, personality, and psychopathology from a social-functional perspective', *Annals of the New York Academy of Sciences* **1000**(1), 222–243.

Kingma, D. P. and Dhariwal, P. (2018), Glow: Generative flow with invertible 1x1 convolutions, *in* 'Proceedings of Advances in Neural Information Processing Systems (NIPS)'.

Klaser, A., Marszałek, M. and Schmid, C. (2008), A spatio-temporal descriptor based on 3d-gradients, *in* 'Proceedings of British Machine and Vision Conference (BMVC)'.

Kleinsmith, A. and Bianchi-Berthouze, N. (2007), Recognizing affective dimensions from body posture, *in* 'Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)'.

Kleinsmith, A. and Bianchi-Berthouze, N. (2013), 'Affective body expression perception and recognition: A survey', *IEEE Transactions on Affective Computing* **4**(1), 15–33.

Knuuttila, S. (2018), 'Medieval theories of the emotions'.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I. (2012), 'Deap: A database for emotion analysis; using physiological signals', *IEEE Transactions on Affective Computing* **3**(1), 18–31.

Koelstra, S. and Patras, I. (2013), 'Fusion of facial expressions and eeg for implicit affective tagging', *Image and Vision Computing* **31**(2), 164–174.

Kollias, D., Nicolaou, M. A., Kotsia, I., Zhao, G. and Zafeiriou, S. (2017), Recognition of affect in the wild using deep neural networks, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)'.

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I. and Zafeiriou, S. (2018), 'Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond', *arXiv preprint arXiv:1804.10938* .

Kotsia, I. and Pitas, I. (2006), ‘Facial expression recognition in image sequences using geometric deformation features and support vector machines’, *IEEE transactions on Image Processing* **16**(1), 172–187.

Kotsia, I. and Pitas, I. (2007), ‘Facial expression recognition in image sequences using geometric deformation features and support vector machines’, *IEEE Transactions on Image Processing* **16**(1), 172–187.

Kuppens, P., Tuerlinckx, F., Russell, J. A. and Barrett, L. F. (2013), ‘The relation between valence and arousal in subjective experience.’, *Psychological Bulletin* **139**(4), 917–940.

Lai, C. and Murray, G. (2018), Predicting group satisfaction in meeting discussions, *in* ‘Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data’.

Lan, T., Sigal, L. and Mori, G. (2012), Social roles in hierarchical models for human activity recognition, *in* ‘Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)’.

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N. and Mori, G. (2012), ‘Discriminative latent models for recognizing contextual group activities’, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **34**(8), 1549–1562.

Lehmann-Willenbrock, N., Hung, H. and Keyton, J. (2017), ‘New frontiers in analyzing dynamic group interactions: Bridging social and computer science’, *Small group research* **48**(5), 519–531.

Leite, I., McCoy, M., Ullman, D., Salomons, N. and Scassellati, B. (2015), Comparing models of disengagement in individual and group interactions, *in* ‘Proceedings of ACM/IEEE International Conference on Human-Robot Interaction’.

Li, J., Roy, S., Feng, J. and Sim, T. (2016), Happiness level prediction with sequential inputs via multiple regressions, *in* ‘Proceedings of ACM International Conference on Multimodal Interaction (ICMI)’.

Li, S. and Deng, W. (2018), 'Deep facial expression recognition: A survey', *arXiv preprint arXiv:1804.08348* .

Lucas, B. D., Kanade, T. et al. (1981), An iterative image registration technique with an application to stereo vision, *in* 'Proceedings of International Joint Conference on Artificial Intelligence'.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. (2010), The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)'.

Lyons, M. J., Budynek, J. and Akamatsu, S. (1999), 'Automatic classification of single facial images', *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **21**(12), 1357–1362.

Mackie, D. M. and Smith, E. R. (2017), 'Group-based emotion in group processes and intergroup relations', *Group Processes & Intergroup Relations* **20**(5), 658–668.

Matsumoto, D. (1989), 'Cultural influences on the perception of emotion', *Journal of Cross-Cultural Psychology* **20**(1), 92–105.

Matsumoto, D. (1991), 'Cultural influences on facial expressions of emotion', *Southern Journal of Communication* **56**(2), 128–137.

McColl-Kennedy, J. R., Patterson, P. G., Smith, A. K. and Brady, M. K. (2009), 'Customer rage episodes: emotions, expressions and behaviors', *Journal of Retailing* **85**(2), 222–237.

McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J. and Picard, R. (2013), Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)'.

Mendolia, M. (2007), 'Explicit use of categorical and dimensional strategies to decode facial expressions of emotion', *Journal of Nonverbal Behavior* **31**(1), 57–75.

Menges, J. I. and Kilduff, M. (2015), 'Group emotions: Cutting the gordian knots concerning terms, levels of analysis, and processes', *The Academy of Management Annals* **9**(1), 845–928.

Miller, G. A. (1995), 'Wordnet: a lexical database for english', *Communications of the ACM* **38**(11), 39–41.

Mioranda-Correa, J. A. and Patras, I. (2018), A multi-task cascaded network for prediction of affect, personality, mood and social context using eeg signals, *in* 'Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)'.

Mollahosseini, A., Hasani, B. and Mahoor, M. H. (2017), 'Affectnet: A database for facial expression, valence, and arousal computing in the wild', *arXiv preprint arXiv:1708.03985* .

Moreland, R. L. (2010), 'Are dyads really groups?', *Small Group Research* **41**(2), 251–267.

Morency, L.-P. (2013), 'The role of context in affective behavior understanding', *Social Emotions in Nature and Artifact* **2**, 8–27.

Mou, W., Celiktutan, O. and Gunes, H. (2015), Group-level arousal and valence recognition in static images: Face, body and context, *in* 'Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)'.

Mou, W., Gunes, H. and Patras, I. (2016), Automatic recognition of emotions and membership in group videos, *in* 'Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)'.

Mou, W., Gunes, H. and Patras, I. (2019*a*), 'Alone vs in-a-group: A multi-modal framework for au-tomatic affect recognition', *ACM Transactions on Multimedia Computing,Communications, and Applications* **15**(47), 1–23.

Mou, W., Gunes, H. and Patras, I. (2019*b*), Your fellows matter: Affect analysis across subjects in group videos, *in* 'Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)'.

Mou, W., Tzelepis, C., Mezaris, V., Gunes, H. and Patras, I. (2018), 'A deep generic to specific recognition model for group membership analysis using non-verbal cues', *Image and Vision Computing* **81**, 42–50.

Niu, X., Han, H., Zeng, J., Sun, X., Shan, S., Huang, Y., Yang, S. and Chen, X. (2018), Automatic engagement prediction with gap feature, *in* 'Proceedings of ACM International Conference on Multimodal Interaction (ICMI)'.

Ojala, T., Pietikäinen, M. and Harwood, D. (1996), 'A comparative study of texture measures with classification based on featured distributions', *Pattern recognition* **29**(1), 51–59.

Ojala, T., Pietikäinen, M. and Mäenpää, T. (2002), 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns', *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **24**(7), 971–987.

Ojansivu, V. and Heikkilä, J. (2008), Blur insensitive texture classification using local phase quantization, *in* 'Proceedings of International Conference on Image and Signal Processing (ICIP)'.

Otsuka, T. and Ohya, J. (1997), Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences, *in* 'Proceedings of International Conference on Image Processing'.

Pai, H.-H., Sears, D. A. and Maeda, Y. (2015), 'Effects of small-group learning on transfer: A meta-analysis', *Educational psychology review* **27**(1), 79–102.

Pantic, M. and Patras, I. (2006), 'Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences', *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **36**(2), 433–449.

Pantic, M., Valstar, M., Rademaker, R. and Maat, L. (2005), Web-based database for facial expression analysis, *in* 'IEEE International Conference on Multimedia and Expo (ICME)'.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A. (2017), 'Automatic differentiation in pytorch'.

Piana, S., Staglianò, A., Camurri, A. and Odone, F. (2013), A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition, *in* 'Proceedings of International Workshop on Intelligent Digital Games for Empowerment and Inclusion'.

Pianesi, F., Zancanaro, M., Lepri, B. and Cappelletti, A. (2007), 'A multimodal annotated corpus of consensus decision making meetings', *Language Resources and Evaluation* **41**(3-4), 409–429.

Picard, R. W. (1995), 'Affective computing'.

Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017), 'A review of affective computing: From unimodal analysis to multimodal fusion', *Information Fusion* **37**, 98–125.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R. and Pantic, M. (2015), Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data, *in* 'Proceedings of International Workshop on Audio/Visual Emotion Challenge'.

Ringeval, F., Sonderegger, A., Sauer, J. and Lalanne, D. (2013), Introducing the recola multimodal corpus of remote collaborative and affective interactions, *in* 'Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)'.

Russell, J. A. (1980), 'A circumplex model of affect.', *Journal of personality and social psychology* **39**(6), 1161.

Sanchez-Cortes, D., Aran, O., Mast, M. S. and Gatica-Perez, D. (2012), 'A nonverbal behavior approach to identify emergent leaders in small groups', *IEEE Transactions on Multimedia* **14**(3), 816–832.

Sánchez, J., Perronnin, F., Mensink, T. and Verbeek, J. (2013), 'Image classification with the fisher vector: Theory and practice', *International Journal of Computer Vision (IJCV)* **105**(3), 222–245.

Sandelands, L. and Clair, L. S. (1993), 'Toward an empirical concept of group', *Journal for the Theory of Social Behaviour* **23**(4), 423–458.

Sapru, A. and Bourlard, H. (2015), 'Automatic recognition of emergent social roles in small group interactions', *IEEE Transactions on Multimedia* **17**(5), 746–760.

Sariyanidi, E., Gunes, H. and Cavallaro, A. (2015), 'Automatic analysis of facial affect: A survey of registration, representation, and recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **37**(6), 1113–1133.

Sariyanidi, E., Gunes, H., Gökmen, M. and Cavallaro, A. (2013), Local Zernike Moment representation for facial affect recognition, *in* 'Proceedings of British Machine and Vision Conference (BMVC)'.

Savran, A., Cao, H., Shah, M., Nenkova, A. and Verma, R. (2012), Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering, *in* 'Proceedings of ACM International Conference on Multimodal Interaction (ICMI)'.

Scherer, K. R., Banse, R., Wallbott, H. G. and Goldbeck, T. (1991), 'Vocal cues in emotion encoding and decoding', *Motivation and emotion* **15**(2), 123–148.

Shalev-Shwartz, S., Singer, Y. and Srebro, N. (2007), Pegasos: Primal estimated sub-gradient solver for SVM, *in* 'Proceedings of International Conference on Machine Learning (ICML)'.

Shan, C., Gong, S. and McOwan, P. W. (2009), 'Facial expression recognition based on local binary patterns: A comprehensive study', *Image and Vision Computing* **27**(6), 803–816.

Shan, C. and Gritti, T. (2008), Learning discriminative lbp-histogram bins for facial expression recognition., *in* 'Proceedings of British Machine and Vision Conference (BMVC)'.

Shrader, A. D. (2015), 'Phd thesis: A comparison of audience response to live and recorded theatre performances', *Marietta College* .

Sidavong, L., Lal, S. and Sztynda, T. (2019), Spontaneous facial expression analysis using optical flow technique, *in* 'Modern Sensing Technologies', pp. 83–101.

Sikka, K., Dhall, A. and Bartlett, M. (2013), Weakly supervised pain localization using multiple instance learning, *in* 'Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG)'.

Sintsovaa, V. and Musata, C. (2013), Fine-grained emotion recognition in olympic tweets based on human computation, *in* 'Proceedings of Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis'.

Sneddon, I., McRorie, M., McKeown, G. and Hanratty, J. (2011), 'The belfast induced natural emotion database', *IEEE Transactions on Affective Computing* **3**(1), 32–41.

Sneddon, I., McRorie, M., McKeown, G. and Hanratty, J. (2012), 'The belfast induced natural emotion database', *IEEE Transactions on Affective Computing* **3**(1), 32–41.

Spence, S. (1995), 'Descartes' error: Emotion, reason and the human brain', *British Medical Journal* **310**(6988), 1213.

Swarbrick, D., Bosnyak, D., Livingstone, S. R., Bansal, J., Marsh-Rollo, S., Woolhouse, M. H. and Trainor, L. J. (2019), 'How live music moves us: head movement differences in audiences to live versus recorded music', *Frontiers in psychology* **9**, 2682.

Tan, L., Zhang, K., Wang, K., Zeng, X., Peng, X. and Qiao, Y. (2017), Group emotion recognition with individual facial emotion cnns and global image based cnns, *in* 'Proceedings of ACM International Conference on Multimodal Interaction (ICMI)'.

Theano Development Team (2016), 'Theano: A Python framework for fast computation of mathematical expressions', *arXiv preprint arXiv:1605.02688* .

Tian, Y.-l., Kanade, T. and Cohn, J. F. (2001), 'Recognizing action units for facial expression analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **23**(2), 97–115.

Valstar, M. and Pantic, M. (2006), Fully automatic facial action unit detection and temporal analysis, *in* 'Proceedings of International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)'.

Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M. and Murino, V. (2016), 'Detecting conversational groups in images and sequences: A robust game-theoretic approach', *Computer Vision and Image Understanding* **143**, 11–24.

Vlachostergiou, A., Caridakis, G. and Kollias, S. (2014), Context in affective multiparty and multimodal interaction: why, which, how and where?, *in* 'Proceedings of ACM Workshop on Understanding and Modeling Multiparty, Multimodal Interactions'.

Wang, H., Kläser, A., Schmid, C. and Liu, C.-L. (2011), Action recognition by dense trajectories, *in* 'Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)'.

Wang, H., Kläser, A., Schmid, C. and Liu, C.-L. (2013), 'Dense trajectories and motion boundary descriptors for action recognition', *International Journal of Computer Vision (IJCV)* **103**(1), 60–79.

Wang, H. and Schmid, C. (2013), Action recognition with improved trajectories, *in* 'Proceedings of IEEE International Conference on Computer Vision (ICCV)'.

Williams, K. D. (2010), 'Dyads can be groups (and often are)', *Small Group Research* **41**(2), 268–274.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B. and Rigoll, G. (2013), 'Lstm-modeling of continuous emotions in an audiovisual affect recognition framework', *Image and Vision Computing* **31**(2), 153–163.

Xiong, X. and De la Torre, F. (2013), Supervised descent method and its applications to face alignment, *in* 'Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)'.

Yang, H., Ciftci, U. and Yin, L. (2018), Facial expression recognition by de-expression residue learning, *in* ‘Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)’.

Yang, J., Wang, K., Peng, X. and Qiao, Y. (2018), Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction, *in* ‘Proceedings of the 2018 on International Conference on Multimodal Interaction (ICMI)’.

Yang, S. and Bhanu, B. (2011), Facial expression recognition using emotion avatar image, *in* ‘Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG)’.

Zafeiriou, S., Papaioannou, A., Kotsia, I., Nicolaou, M. A., Zhao, G., Antonakos, E., Snape, P., Trigeorgis, G. and Zafeiriou, S. (2016), Facial affect “in-the-wild”: A survey and a new database, *in* ‘Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)’.

Zancanaro, M., Lepri, B. and Pianesi, F. (2006), Automatic detection of group functional roles in face to face interactions, *in* ‘Proceedings of ACM International Conference on Multimodal Interfaces (ICMI)’.

Zeng, Z., Pantic, M., Roisman, G. I. and Huang, T. S. (2009), ‘A survey of affect recognition methods: Audio, visual, and spontaneous expressions’, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **31**(1), 39–58.

Zhang, L. and Hung, H. (2016), Beyond f-formations: Determining social involvement in free standing conversing groups from static images, *in* ‘Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)’.

Zhao, G. and Pietikainen, M. (2007), ‘Dynamic texture recognition using local binary patterns with an application to facial expressions’, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **29**, 915–928.

Zhao, G. and Pietikäinen, M. (2009), ‘Boosted multi-resolution spatiotemporal descriptors for facial expression recognition’, *Pattern Recognition Letters* **30**(12), 1117–1127.

Zhu, Y., Heynderickx, I. and Redi, J. A. (2014), Alone or together: measuring users’ viewing experience in different social contexts, *in* ‘Human Vision and Electronic Imaging XIX’, Vol. 9014, pp. 90–101.