Investigating Multi-Modal Features for Continuous Affect Recognition Using Visual Sensing

Haolin Wei, B.Sc. (hons)

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Electronic Engineering

Supervisors:

Prof. Noel E. O'Connor

Dr. David S. Monaghan

January 5, 2018

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No: 57106126

Date:

Abstract

Emotion plays an essential role in human cognition, perception and rational decision-making. In the information age, people spend more time then ever before interacting with computers, however current technologies such as Artificial Intelligence (AI) and Human-Computer Interaction (HCI) have largely ignored the implicit information of a user's emotional state leading to an often frustrating and cold user experience. To bridge this gap between human and computer, the field of affective computing has become a popular research topic. Affective computing is an interdisciplinary field encompassing computer, social, cognitive, psychology and neural science. This thesis focuses on human affect recognition, which is one of the most commonly investigated areas in affective computing. Although from a psychology point of view, emotion is usually defined differently from affect, for this thesis the terms *emotion*, *affect*, *emotional state* and *affective state* are used interchangeably.

Both visual and vocal cues have been used in previous research to recognise a human's affective states. For visual cues, information from the face is often used. Although these systems achieved good performance under laboratory settings, it has proved a challenging task to translate these to unconstrained environments due to variations in head pose and lighting conditions. Since a human face is a three-dimensional (3D) object whose 2D projection is sensitive to the aforementioned variations, recent trends have shifted towards using 3D facial information to improve the accuracy and robustness of the systems. However these systems are still focused on recognising deliberately displayed affective states, mainly prototypical expressions of six basic emotions (happiness, sadness, fear, anger, surprise and disgust). To our best knowledge, no research has been conducted towards continuous recognition of spontaneous affective states using 3D facial information.

The main goal of this thesis is to investigate the use of 2D (colour) and 3D (depth) facial information to recognise spontaneous affective states continuously. Due to a lack of an existing continuous annotated spontaneous data set, which contains both colour and depth information, such a data set was created. To better understand the processes in affect recognition and to compare results of the proposed methods, a baseline system was implemented. Then the use of colour and depth information for affect recognition were examined separately. For colour information, an investigation was carried out to explore the performance of various state-of-art 2D facial features using different publicly available data sets as well as the captured data set. Experiments were also carried out to study if it is possible to predict a human's affective state using 2D features extracted from individual facial parts (E.g. eyes and mouth). For depth information, a number of histogram based features were used and their performance was evaluated. Finally a multi-modal affect recognition framework utilising both colour and depth information is proposed and its performance was evaluated using the captured data set.

Acknowledgements

Firstly, I would like to greatly thank my supervisors Prof. Noel E. O'Connor, Dr Gabriel Miro Muntean, Dr Patricia Scanlan and Dr David Monaghan for all their guidance and support during my time pursing this interesting yet challenging research project.

Thanks to all the current and past members of Insight who offered advice and support. I would particularly like to thank Dr Kevin McGuinness, Dr Philip Kelly, Marc Gowing, Edmond Mitchell, Dr Dian Zhang, Jinlin Guo and Dr Feiyan Hu for their help on every step of my research. Lastly, I would like to thank my wife, Lingge Li and my parents for their understanding and support.

This work is co-funded by Bell Labs Ireland and the Irish Research Council under the Enterprise Partnership scheme. The research is also supported in part by the European Commission under the Contract FP7-ICT-287723 REVERIE and European Union's Horizon 2020 Framework Programme under Grant Agreement no. 643491.

Contents

List of Figures

List of Tables

List of Publications

- Kuijk, F., Apostolakis, K.C., Daras, P., Ravenet, B., Wei, H., & Monaghan, D.S.(2015 June). Autonomous agents and avatars in REVERIE's virtual environment. In *Proceedings of the the 20th International Conference on 3D Web Technology* (pp. 279-287). ACM.

- Wei, H., Monaghan, D.S., OConnor, N.E., & Scanlon, P. (2014, September). A new multi-modal dataset for human affect analysis. In *International Workshop on Human Behavior Understanding* (pp. 42-51). Springer International Publishing.

- OConnor, N.E., Alexiadis, D., Apostolakis, K., Daras, P., Izquierdo, E., Li, Y., Monaghan, D.S., Rivera, F., Stevens, C., Van Broeck, S. and Wall, J., & Wei, H. (2014, January). Tools for user interaction in immersive environments. In *International Conference on Multimedia Modeling* (pp. 382-385). Springer International Publishing.

- Li, Y., Wei, H., Monaghan, D.S., & OConnor, N.E. (2014, January). A low-cost head and eye tracking system for realistic eye movements in virtual avatars. In *International Conference on Multimedia Modeling* (pp. 461-472). Springer International Publishing.

- Wei, H., Scanlon, P., Li, Y., Monaghan, D.S., & O'Connor, N.E. (2013, July). Real-time head nod and shake detection for continuous human affect recogni-

tion. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)* (pp. 1-4). IEEE.

Chapter 1

Introduction

1.1 Motivation

In most parts of the world digital devices such as computers and mobile phones have become ubiquitous. We use them in work, communication, shopping and entertainment. This growing ubiquity has led to a focus on human-centred design. One of the key components of human-centred design is human computer interaction (HCI). Recent years have witnessed a trend away from traditional HCI such as mouse and keyboard to a more user-centred design such as the use of hand gestures and voice commands. However, these designs still ignore the user's emotional states, which is a fundamental component of human-to-human communication. As a result, such interaction not only filters out a large amount of information available in the interaction process, but also means that the interaction experience is frequently perceived as cold, impersonal and frustrating (Zeng *et al.*, 2009).

In addition to HCI scenarios, systems that can sense people's emotional states are also potentially beneficial for customer services, call centres, e-learning, intelligent autonomous vehicles, games and other entertainment. For example: an autonomous monitoring system in an e-learning environment could provide live feedback to the lecturer on the status of the students; an automatic call centre could decide when to pass the customer to human operators (Lee and Narayanan, 2005) and an intelligent vehicle could monitor the tiredness or mood of the driver to potentially avoid collisions (Ji *et al.*, 2006). Other important applications of automatic affect recognition are in more traditional/theoretical affect-related research such as psychology, behavioural science and neuroscience, where such systems could improve the reliability of measurements in subjective human trials and speed up currently time-consuming manual processing of human affective behaviour data (Ekman and Rosenberg, 1997).

Given the practical and theoretical importance of this field, a significant body of research in the past 40 years has been conducted towards automatic affect recognition. Williams and Stevens (1972) presented the first attempt to identify and measure parameters from speech signals that reflect a speaker's emotional states. Suwa *et al.* (1978) showed an early attempt to analyse facial expressions automatically by tracking the motion of twenty identified points on the face. With advances in machine learning, Kobayashi and Hara (1991) proposed to recognise the six basic emotions (happiness, sadness, fear, disgust, anger and surprise) from face images using a neural network. The study carried out by Dellaert *et al.* (1996) compared multiple pattern recognition techniques to recognise a subset of the basic emotions (happiness, sadness, anger and fear) from speech. Motivated by the limitations of using only visual or vocal cues, Chen *et al.* (1998) demonstrated one of the first attempts at using both audio and visual modalities for emotion recognition and showed an improvement in overall recognition accuracy of the six basic emotions. Traditionally, facial information used in affect recognition has been based on 2D data (colour pixel information) which is usually limited by head pose and illumination changes. In order to tackle these problems 3D data can be used. An early attempt at recognising the six basic emotions from static 3D facial data was carried out by Wang *et al.* (2006), where the authors showed improved results compared to the use of static 2D facial data. More recently, Tsalakanidou and Malassiotis (2009) presented a system that uses dynamic 3D facial data for affect recognition.

Although a seminal study by Ambady and Rosenthal (1992) suggested that visual cues from the face and body are both important for judging people's emotional state, the existing literature on automatic affect recognition did not focus on the information carried by the body until 2004 (Camurri *et al.*, 2004). Since then, a number of researchers have attempted to combine facial expressions and body gestures for affect recognition (Gunes and Piccardi, 2007; Karpouzis *et al.*, 2007). Interest in detecting emotions from physiological signals emerged from the well known work of Picard *et al.* (2001) where the authors use four physiological sensors including electromyogram (EMG), photoplethysmography (PPG), galvanic skin response (GSR) and respiration sensors to recognise eight emotional states (neutral, anger, hatred, grief, platonic love, romantic love, joy and reverence). Since then a number of approaches have also been proposed for other types of sensors such as Electroencephalographic (EEG) (Takahashi, 2004; Nakasone *et al.*, 2005) and thermal infrared cameras (Tsiamyrtzis *et al.*, 2007).

As can be seen, most of early research on automatic affect recognition has focused mainly on recognising discrete emotional states, however, a number of researchers have shown that a single label (or small number of discrete classes) may not reflect the subtle and complex affective states that occur in everyday interactions (Russel, 1980). Hence, the recent trend has shifted towards the use of a dimensional description of human affect, where an affective state is represented by a number of latent dimensions (Russel, 1980). Two of the most commonly used dimensions are valence and arousal, where valence reflects how a person feels, from positive (e.g. happy and joy) to negative (e.g. anger and fear) and arousal reflects how likely a person is to take an action, from active to passive. Considering automatic dimensional affect recognition, early attempts usually simplify the problem to a two-class (positive-negative or active-passive) or a four-class (positive-active, positive-passive, negative-active, negative-passive) valence-arousal related classification problem (Fragopanagos and Taylor, 2005; Caridakis *et al.*, 2008; Glowinski *et al.*, 2008; Schuller *et al.*, 2011). Since 2011, a number of researchers have started to treat the automatic dimensional affect recognition problem as a regression problem. For instance, Nicolaou *et al.* (2011) presented one of the first attempts to continuously recognise spontaneous affect in the valance-arousal parameter space using both visual and vocal cues. The current focus in automatic affect recognition research is to continuously recognise the spontaneous affective state using multi-modal cues in this dimensional space (Schuller *et al.*, 2012; Valstar *et al.*, 2013, 2014; Ringeval *et al.*, 2015b).

While the current approaches in continuous spontaneous affect recognition have proven to be highly successful using data captured in controlled laboratory settings, little work have been done on using data from an unconstrained environment (e.g. with diverse illumination conditions and backgrounds). This forms the motivation in this thesis for the investigation of different approaches that meet the requirements of real-world applications.

Firstly, when using descriptors to represent different affective states, these descriptors should be invariant to any change of the captured subject. For example when a person's face is only partially visible due to being covered by hair or looking away from the camera, the affect recognition system should still give a correct prediction of a person's affective state.

Secondly, due to the variety of environments that might be encountered in practice, there are situations where a colour image is not clearly visible (e.g. under extreme low-light conditions). The affect recognition system should be able to adapt to the environment and use additional modalities to provide a correct prediction.

These two requirements motivated the main goal of this thesis, which is to investigate how visual cues could be used for continuous spontaneous affect recognition. The visual cues consist of colour information which can be captured by a traditional camera, and depth information which can be obtained using a depth sensor. This thesis focuses on the visual cues from the human face since the face is a crucial modality in conveying human expressions. Facial expressions could indicate a person's affective state, intentions and ultimately, elicit other people's response. For instance, Segal (2008) suggested that it is possible to infer other people's affective state just by looking at the individual's face, without any complementary information such as voice or gesture, indicating that the face could be the most effective communication tool.

1.2 Research Objectives

There are several research objectives identified in this thesis when using visual cues for continuous spontaneous affect recognition. The initial research objective is to thoroughly investigate the performance of popular low-level appearance features for continuous affect recognition. In particular, this thesis focuses on histogram-based features since they are one of the most commonly used appearance features for continuous affect recognition. To accomplish this objective, experiments are carried out to select the best configurations for different histogram-based features using a publicly available dataset. Then using the best configurations, different histogram-based features are extracted and evaluated on other datasets in order to study which features give the best performance for continuous affect recognition across different capture settings.

The next research objective is to study if individual facial parts such as mouth, and eyes could be used to recognise affective state. The motivation behind this research objective is that most of the existing approaches use features extracted based on the entire face image. However one limitation of this approach is that it might not generalise well for situations such as partial occlusions. If it is possible to recognise affective state from individual facial parts, this should increase the robustness of the system to partial occlusions.

Another research objective is to investigate if features extracted from low cost depth sensors could be used for continuous affect recognition. To accomplish this objective, a face detection algorithm based on a depth image is first proposed. Then the histogram-based features are applied to the data captured by the depth sensor and their performance is evaluated.

The final research objective is to investigate if the use of both colour and depth features lead to a significant increase in performance in affect recognition. With features from different modalities, the issue of fusion is a crucial consideration. In this context both feature-level fusion and decision-level fusion need to be investigated.

1.3 Research Contributions

The main scientific contributions provided by this thesis are summarised as:

- Designing and developing a novel data capture platform for collecting synchronised video, depth and audio streams.

- To the best of my knowledge, I captured one of the first spontaneous and continuously annotated multi-modal dataset based on human interaction during a debate.

- Thoroughly investigating the histogram-based appearance features for continuous affect recognition.

- Proposing automatically generated appearance features from facial parts for continuous affect recognition.

- Extending the low level appearance features to the depth modality for continuous affect recognition.

- Developing a multi-modal framework for continuous affect recognition using colour and depth information.

1.4 Thesis Outline

The remainder of this thesis is structured as follows:

In Chapter 2, the relevant technical background and in particular the limitations of the current affect recognition systems are reviewed. First, different affect models used to represent a human's affective state are introduced. Next, an overview of the existing datasets used for affect recognition including how these datasets were captured and annotated are examined. Then different features from video and depth modalities used for affect recognition are explored before reviewing machine learning techniques. Finally, the performance evaluation metrics used for affect recognition are discussed.

In Chapter 3, the steps of designing and developing a multi-modal data capture platform that could collect synchronised colour, depth and audio streams are first described. Then a detailed description is given on how this platform can be used to construct a three-way debate affect dataset. The processes for segmentation and annotation of the dataset are then presented, and finally a statistical analysis of the dataset is performed to demonstrate its value as a research tool.

In Chapter 4, a thorough investigation of which histogram based features from the colour modality give the best affect recognition results is carried out. In particular, the first set of experiments concentrate on evaluating the performance of hand-crafted low-level appearance features extracted from an entire face region, while the second set of experiments focus on automatically generated features from individual facial parts. These experiments are carried out using different datasets including two publicly available datasets, along with the dataset captured in the previous Chapter.

In Chapter 5, a face detector that using depth images is first proposed. Then the depth face detector is used to extract the face region from the depth data captured in Chapter 3. The histogram-based features are applied to the depth data, and their performance is examined. Various experiments are performed to study how predictions from the video and depth modalities could be fused. Specifically, early fusion and late fusion methods are evaluated. Finally, a multi-modal affect recognition framework which can be used in real world settings is proposed.

In Chapter 6, outcomes of the research work to date are presented and further work directions are proposed.

Chapter 2

Related Work and Background

2.1 Introduction

In this chapter an overview of the background literature essential to understanding the core work carried out in this thesis are given. In section 2.2, various computational models for human emotional processes are introduced. Their advantages and disadvantages are discussed. In section 2.3, a review of the currently available datasets for affect recognition is given. In section 2.4, the current state-of-the-art techniques used in affect recognition are explored. First, different features used for 2D and 3D signals are examined, followed by a review of the machine learning techniques that will later be employed for the work carried out in this thesis. Finally, the performance evaluation metrics commonly used in affect recognition are explained.

2.2 Emotion Models

There has been a long debate on the nature and the causal generation of emotion by psychologists since the latter third of the nineteenth century when psychology began to form as an independent academic discipline (Calvo *et al.*, 2014). Driven by the needs for basic emotion and cognition research, various computational emotion models have been proposed along with the theories. These computational models have bridged the gap between psychology and computer science. Currently there are three commonly used models in affective computing. This section briefly introduces these models. A more detailed discussion on each of these models can be found in Fox (2008) and Scherer *et al.* (2010).

2.2.1 Categorical Model

Influenced by Darwin's evolutionary view of emotions (Darwin, 1872), the discrete categorical emotion model suggests that there are a set of basic emotions that drive the motives of human beings (Tomkins, 1962, 1963). Most of the categorical models describe emotions using a list of affective-related keywords. By showing subjects still photographs of acted facial expressions, Ekman (1971) concluded that six basic emotions can be universally recognised. These emotions are happiness, sadness, fear, anger, disgust, and surprise (See Figure 2.1 for an example of each emotion). Sometimes, an additional neutral label is also added resulting in seven basic emotions. Although the number of basic emotions varies from 2 to 18 depending on different theories (James, 1884; Izard, 1971; Frijda and Swagerman, 1987; Ortony and Turner, 1990; Wierzbicka, 1992), there has been considerable agreement on Ekman's six basic emotion categories.

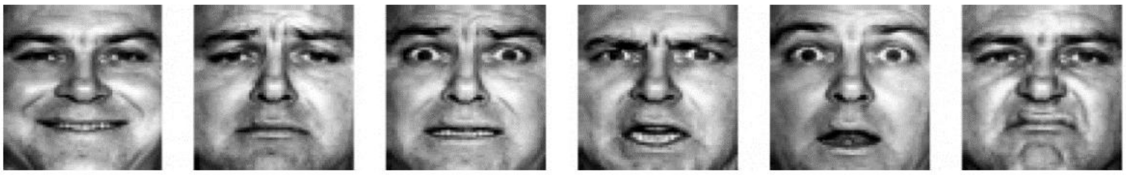


Figure 2.1: Facial expressions of the six basic emotions - happiness, sadness, fear, anger, surprise and disgust (Ekman and Friesen, 1976)

However, a number of psychology researchers suggest that it is necessary to go beyond discrete emotions. For instance, Baron-Cohen *et al.* (2004) argues that it is important to include cognitive mental states such as agreement, interest, thinking and concentrating in addition to the basic emotions, as they occur more often in everyday interaction compared to the basic emotions. To date, the discrete categorical model has been the most commonly adopted approach in automatic affect recognition research. The advantage of the category representation is that people use these words in daily life to describe observed emotions, which makes this scheme intuitive to use and understand. However, there are two main drawbacks of the categorical model: (a) it can not reflect the change in intensity of people's affective state continuously, for example people may have different levels of happiness from pleasure, joy to ecstasy; (b) it fails to express subtle or blended affective states like depression, contempt and embarrassment that could occur in a natural communication setting.

2.2.2 Dimensional Model

The dimensional model is an alternative approach to the categorical description of human affect. It is based on the pioneering work of Wundt (1905). They argued that feelings can be described by the dimensions of *strain-relaxation*, *arousing-inhibition* and *pleasantness-unpleasantness*. Similarly, the study carried out by Osgood *et al.* (1975) on affective meaning suggested that three key dimensions namely *evaluation* (good or bad), *potency* (strong or weak) and *activity* (active or passive) exist in almost every language/culture community. The dimensional model assumes that affective states are not independent from one another; instead, they are related to one another by the most fundamental affective feelings called *core affect* (Russell and Barrett, 1999). Typically, two to four dimensions are usually used to represent core affect. One of the most widely used dimensional models is the valence-arousal model proposed by Russel (1980). The model is usually represented by a circular configuration called *Circumplex of Affect* where each axis indicates one core affect, and different emotional labels could be plotted at various positions on the two-dimensional plane (See Figure 2.2). The valence dimension reflects how people feel, from positive (e.g. happy and joy) to negative (e.g. anger and fear). The arousal dimension reflects how likely a person is to take an action: low arousal indicates less energy to take an action, high arousal indicates more energy to take an action.

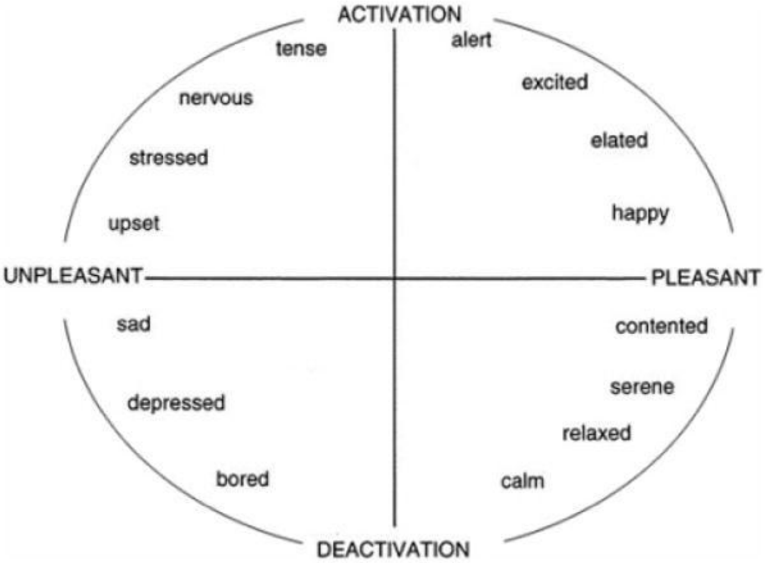


Figure 2.2: A graphical representation of the circumplex model of affect with the horizontal axis representing the valence dimension and the vertical axis representing the arousal dimension (Russel, 1980)

As the two-dimensional model cannot easily differentiate some affective states that share similar values for valence and arousal (e.g. anger and fear, which both result in high arousal and negative valence), a third dimension called dominance could be added. The dominance dimension reflects how much a person feels in control (e.g. anger indicates high dominance and fear indicates low dominance). This model is usually referred to as the *PAD emotion space* for pleasure, arousal and dominance (Mehrabian, 1995) or as *emotional primitives (Espinosa* et al.*, 2010).* The study carried out by Fontaine *et al.* (2007) suggested to also include expectation (the degree of anticipating) as a fourth dimension to distinguish better emotions such as surprise from other affective states. Compared to the categorical model, the dimensional model is able to offer more flexibility when analysing emotions. However, as pointed out by Russel (1980), the dimensional approach is only useful to characterize core affect instead of *complete emotions* since *complete emotions* only fall into certain regions of the space defined by the core affect dimensions. For example when the valence-arousal dimensional model is used, fear and anger could share identical core affects. Even when the four-dimensional model is used, it is still difficult to differentiate emotions such as shame, guilt and embarrassment (Fontaine *et al.*, 2007). Another challenge of using the dimensional approach is the annotation delay when continuously annotating data for scientific experiments. This delay is mainly caused by the reaction time between an annotator perceiving the affective state, then giving the corresponding evaluation scores. The use of such delayed ground truth could lower the performance of the affect recognition system.

2.2.3 Appraisal Model

The appraisal models or componential appraisal models are based on the pioneering work of Arnold (1960) and Lazarus (1966) where emotion is mainly seen as determined by *appraisal*, and it is generated through *continuous*, *recursive subjective evaluation* of both people's potential and the status of their environment. For example, a farmer who sees an approaching bear will react differently from a hunter (e.g. fear vs. excitement) based on the evaluation of individual's potential. The same farmer who sees a bear in a zoo will also react differently to when he sees the bear in the wild, based on the evaluation of the environment (e.g. interest vs. fear). One of the most widely used appraisal models is the OCC model developed by Ortony, Collins and Clore (1988), which describes the cognitive structure of emotions. The OCC model proposed a hierarchy that classifies 22 emotions along with three main branches as shown in Figure 2.3. The three branches are (1) emotions concerning consequences of events (e.g. pleased and displeased), (2) actions of agents (e.g. approving and disapproving), and (3) aspects of objects (e.g. liking and disliking). Each emotion is then treated as a *valenced* (positive or negative) reaction in terms of one of the three main branches as described in Ortony *et al.* (1988). In addition, some subsequent branches could combine together to form compound emotions.

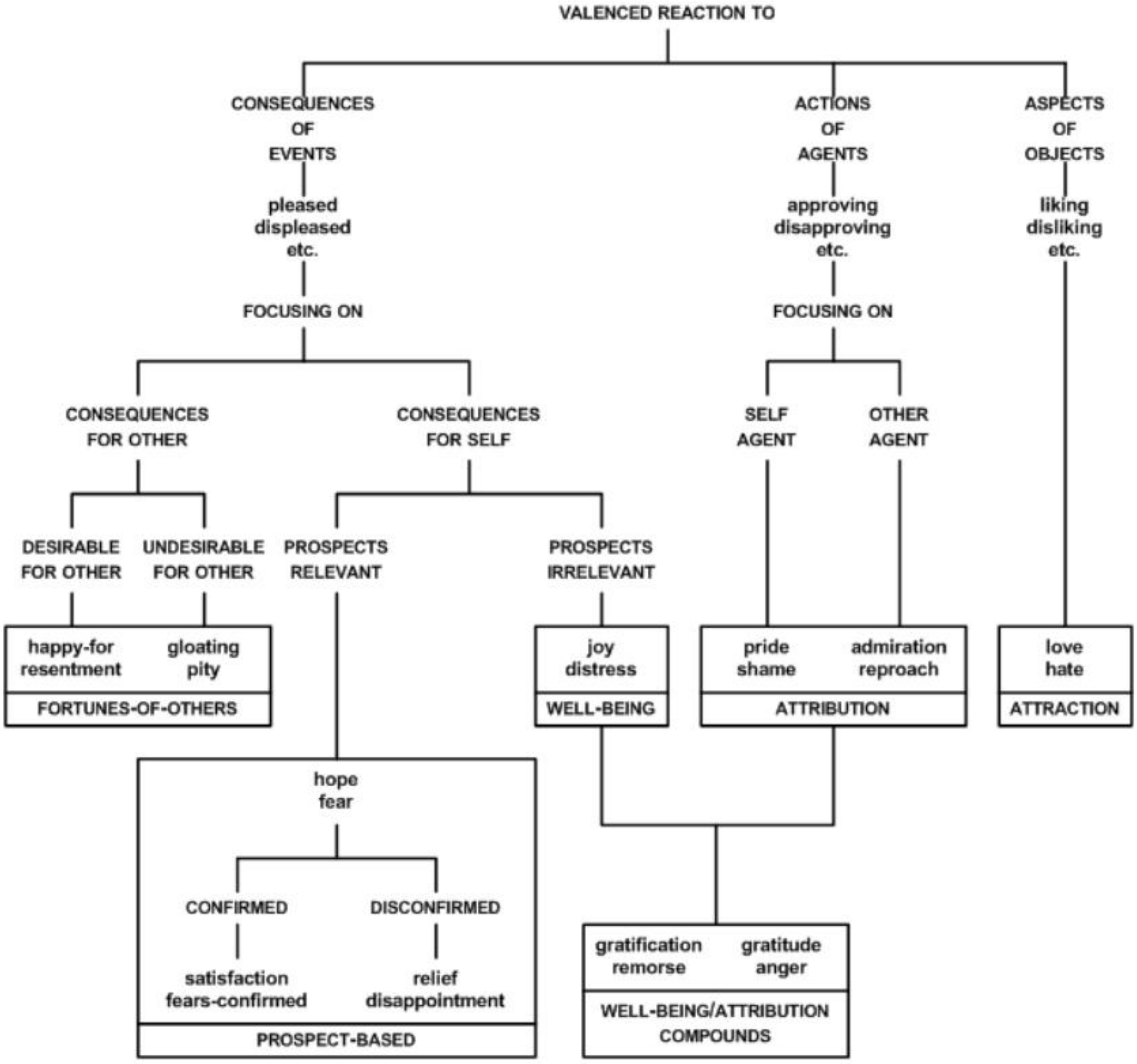


Figure 2.3: Structure of the OCC model (Ortony *et al.*, 1988)

The advantage of the appraisal model is that it does not limit affective states to a fixed number of discrete categories or a set of affect dimensions. In contrast, it focuses on the variability of affective states as a result of change in all relevant contributing factors including *cognition*, *motivation*, *physiological reactions*, *motor expressions* and *feelings* (Gunes, 2010; Calvo *et al.*, 2014). This makes it possible to differentiate and model the *full-blown emotion* space. However, the appraisal model requires *complex, multi-componential and sophisticated measurements of change.* How to use it in automatic affect recognition with such a framework remains an open research question (Gunes, 2010).

2.2.4 Discussion

While the discrete categorical model has been one of the most widely used approaches in affective computing, the recent trend has started to shift towards the use of dimensional models (Schuller *et al.*, 2011, 2012; Valstar *et al.*, 2013; Ringeval *et al.*, 2015b). As suggested by Schuller *et al.* (2012), the dimensional model has shown the ability to encode small differences in affect over time, and to distinguish better between subtly different affective states compared to the limited categorical approach, while at the same time providing an easy-to-implement framework compared to the intractable appraisal model. As a result, the dimensional approach was used for the automatic affect recognition research described in this thesis. In addition, continuously annotated arousal and valence dimensions are chosen for this research due to their widespread use in current affect recognition research (Metallinou *et al.*, 2013; Schuller *et al.*, 2011, 2012; Ringeval *et al.*, 2015b).

2.3 Emotion and Affect Datasets

Affect recognition requires rich sets of labelled (Ringeval *et al.*, 2013) and application specific data (Afzal and Robinson, 2009; Cowie *et al.*, 2010a). Only with such data is it possible to start to train a computer to recognise affect. This section briefly introduce various techniques used to construct an affect dataset, including how affective states can be elicited, what modalities are usually captured and how data is annotated. Finally a comparison of existing datasets is presented.

2.3.1 Eliciting Affective State

As Ringeval *et al.* (2013) suggested, in general, there are three types of interaction behaviour that have been used to elicit human affect. The first is posed behaviour, where the participant is asked to perform a certain affective state such as happy and sad. The second is induced behaviour, where the participant is put in a controlled environment to elicit a certain affective state. The induced affective behaviours are usually captured in two scenarios: Human Computer Interaction (HCI) or the use of a video kiosk. HCI scenarios include Wizard of Oz scenarios (Batliner *et al.*, 2004; Douglas-Cowie *et al.*, 2007) and computer-based dialogue systems (Lee and Narayanan, 2005). In particular, the Wizard of Oz scenario captures participants' reaction while interacting with a computer system that participants believe to be autonomous, but which is actually being operated by an unseen human being. The video kiosk scenario recodes participants' reaction while they are watching emotion-inducing videos. The third type is spontaneous behaviour, which appears in a real-life setting through human-human interactions such as face-to-face interviews (Bartlett *et al.*, 2005), phone conversations (Devillers and Vasilescu, 2004), meetings (Burger *et al.*, 2002) and debates (Grimm *et al.*, 2008).

Among all three types of interaction scenarios, the posed affect is the easiest to design and capture. However, it has been proven that the affective states elicited from a real-life context are more subtle than the posed ones, as in the posed scenario people tend to exaggerate the affective state they are displaying (Gunes and Schuller, 2013). The induced affective state could provide a natural emotional response. However, it is usually not suitable in terms of covering the full range and complexities of affective states, as the interaction is restricted to a specific context (McKeown *et al.*, 2012; Ringeval *et al.*, 2013). Finally the spontaneous affective state is the hardest to capture, as true affective states are relatively *rare, short lived, and filled with subtle context-based changes* (Gunes and Schuller, 2013). Furthermore, informing a participant that they are being recorded could lead to a change in natural behaviours. However, not informing participants that they are being recorded raises potential ethical issues. In order to ethically capture spontaneous affective state, various techniques have been developed. For example, in Zhang *et al.* (2013) the authors use a series of activities such as, listening to a joke or experiencing harsh insults from the experimenter to try to elicit a target emotional state, while the authors in Ringeval *et al.* (2013) use survival task techniques where group discussion is promoted by asking participants to reach a consensus on how to survive in a disaster scenario.

2.3.2 Modalities and Cues

In affective computing, a modality is usually defined as the single independent communication channel of sensory information between a human and a computer (Karray *et al.*, 2008). Based on the nature of different modalities, they are usually divided into three categories: *Visual Modality*, *Audio Modality* and *Biomedical Modality*. An overview of different types of modalities and cues can be seen in Figure 2.4. Whilst this thesis focuses on the visual modality, for completeness, we discuss all three modalities in the following subsections.

2.3.2.1 Visual Modality

The visual modality has been the most widely used modality in the literature for capturing affective states. Usually two types of cues from visual signals are used for automatic affect recognition, namely facial expressions and gestures.

Most of the vision-based affect recognition studies have been focused on facial expression analysis, due to the importance of the face in emotion expression and perception. A facial expression usually refers to one or more motions or positions of the muscles under the skin of the face. There are two main trends in the recent

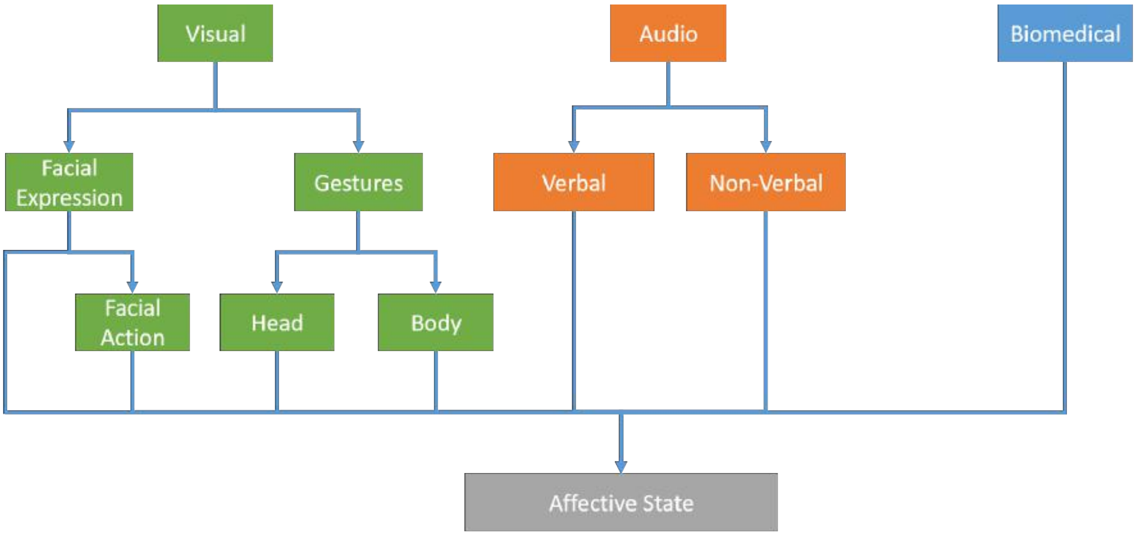


Figure 2.4: Overview of modalities and cues used in affective computing

research on how facial expressions can be used: i) facial expressions are mapped directly to selected emotion models, ii) facial actions are detected first, then mapped to selected emotion models. These two trends are directly derived from the two main approaches used for facial expression measurement in psychological research: message and sign judgement (Cohn, 2006). Message judgement is used to interpret what underlies a displayed facial expression, while sign judgement is used to describe the actual facial behaviour. For example, a wide-eyed expression can be judged as surprise in message judgement terms and in sign judgement terms it could be described as eyebrows raised, eyes widened and mouth open. To label different facial actions, the Facial Action Coding System (FACS) developed by Ekman and Friesen (1976) is widely used. The FACS consists of 44 Action Units (AUs) that in turn represent the movement of individual or groups of facial muscles as shown in Figure 2.5 and Figure 2.6. Compared to the message judgement approach, the sign judgement approach is more robust to context-dependent or culture-specific expressions (Baltrusaitis *et al.*, 2011). However, labelling facial actions requires experienced annotators and is a time-consuming process.

Gestures such as head and body movements are other visual cues that can be

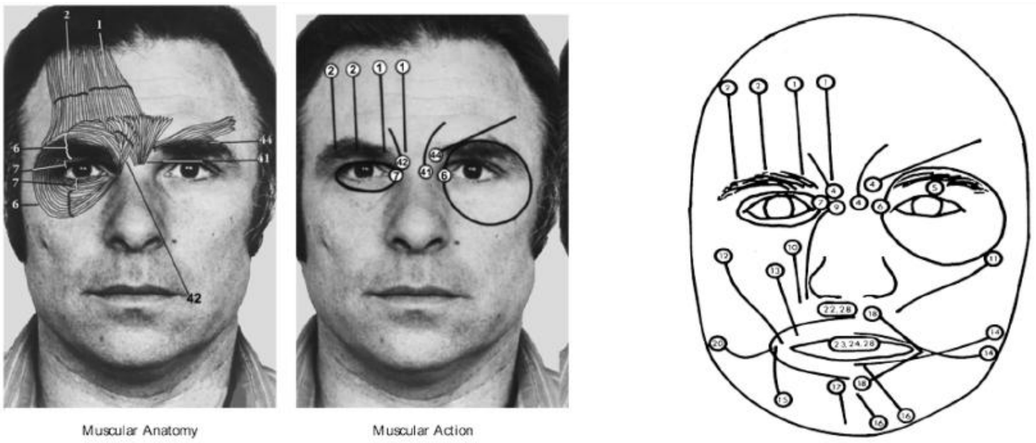


Figure 2.5: Left: Relation between muscular anatomy and muscular action. Right: The AUs of FACS. The small circles represent fixed points towards which skin is pulled along the line during activation while the number in the circle represents the AU. Both images come from Ekman and Friesen (1978)

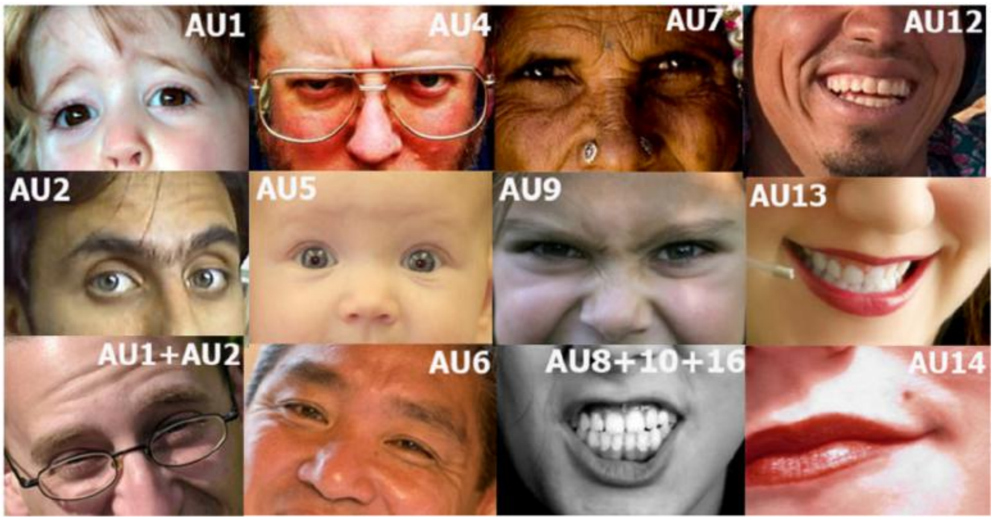


Figure 2.6: Example of facial AUs and their combinations (Pantic and Bartlett, 2007)

used to interpret emotional states. Early research has been focused on mapping body gestures to discrete emotion categories. For instance, a study carried by Darwin (1998) suggested that when people are angry, a number of cues could be observed including i) whole body trembles, ii) head is erect, iii) chest is well expanded, iv) feet are firmly on the ground and v) elbows are squared. The work reported by Wallbott (1998) found that certain distinctive features exist in body movement that can be used to classify specific emotions. For example, erect body posture is rare when experiencing shame, sadness or boredom, lifting shoulders is typical for elated joy and anger, and a head moving downward is most typical when expressing disgust. Coulson (2004) showed that recognition results from body gestures could be as significant as the voice modality and facial expression in some cases. Good recognition results for discrete emotions from body information are also reported by Van den Stock *et al.* (2007). Recently, the trend has shifted towards using gestures to interpret dimensional affective state. By investigating the emotional and communicative significance of head gestures in a naturalistic dataset, Cowie *et al.* (2010b) found that head nods carry informations on both arousal and valence dimensions, while head shakes are good indicators on the arousal dimension. Gunes and Pantic (2010) explored the use of conversational head movement for continuous affect prediction and suggested that head gestures can be used to predict the arousal, valence, dominance and expectation dimensions. Nicolaou *et al.* (2011) on the other hand investigated the use of shoulder gestures for dimensional and continuous affect recognition and found shoulder gestures outperformed vocal cues on valence prediction, and achieved similar performance on the arousal dimension compared to facial cues.

The visual modality is usually captured with a conventional 2D camera. It works by projecting a 3D scene onto a 2D image plane and the captured image is usually referred to as a colour image. However, during this process the distance (range/depth) information is lost. More recently, the trend has shifted to also include 3D cameras to record distance information along with the colour information. The image that contains distance information is often referred as the depth image or range image.

2.3.2.2 Audio Modality

Audio is another commonly used modality in automatic affect recognition research. There are mainly two types of cues from the audio modality that can be used to capture different affective states, namely verbal cues and non-verbal cues.

Verbal cues refer to the linguistic content of the speech. Researchers have shown it is possible to interpret a speaker's affective state through the words he/she used. For instance, Whissell (2009) revised the Dictionary of Affect in Language (DAL) which includes 8,472 words with a 2D rating in the arousal/valence space. Studies carried out by Wöllmer *et al.* (2010) suggested words such as *again*, *angry* and *very* are correlated with the arousal dimension, while *good*, *great* and *totally* are more correlated with the valence dimension.

Non-verbal cues refer to the non-linguistic part of the communication. Research in psychology and psycholinguistics has shown that acoustic and prosodic features can be used to encode the affective states of a speaker. For instance, acoustic parameters such as mean of the fundamental frequency (F0), mean intensity, speech range and high-frequency energy are positively correlated with the arousal dimension (Huttar, 1968; Scherer and Oshinsky, 1977), while prosodic features such as pause duration, pausing and breathing rate are indicative of excitement (Trouvain and Barry, 2000). Other non-verbal cues such as sighs and gasps can also convey emotion information. Sighs usually arise from a negative affective state, such as boredom or dissatisfaction, or at the end of some negative situation as relief, while gasps could occur from an emotion of surprise, shock or disgust (Nicolaou, 2009).

Research has shown that non-verbal cues are good at predicting the arousal dimension, while verbal cues are good at predicting the valence dimension. However, current research in automatic affect recognition mainly focuses on using non-verbal cues, since verbal cues are usually language dependent and it is difficult to anticipate a person's word choice for expressing different emotional states.

The audio modality is usually captured with a microphone. When capturing multi-people interactions, an additional headphone is usually worn by each participant to prevent the microphone capturing other participants' voice during conversations.

2.3.2.3 Biomedical Modality

Biomedical signals are multichannel recordings of physiological activities of organisms. A number of studies have shown that biomedical signals could reflect people's affective states. For example, Galvanic Skin Response (GSR) provides a measurement of skin conductance (See Figure 2.7). It is positively correlated with a person's overall arousal or stress level (Ekman *et al.*, 1983; Levenson, 1992) and it can be used to distinguish emotional states such as fear and anger (Ax, 1953). Heart rate sensors measure heart beat per minute, which increases with negative emotions such as anxiety and fear (Chanel *et al.*, 2007). Respiration rate describes how deep and fast a person is breathing. Irregular and quick breathing usually indicates more aroused emotions such as anger or fear (Chanel *et al.*, 2007). Electromyography (EMG) measures muscle electrical activity when at rest and during contraction (See Figure 2.8). It is correlated with negatively valenced emotions (Lang, 1995).

Compared to sensors used to capture visual and audio modalities, biomedical sensors are usually perceived as being inconvenient to set up and wear, since these sensors usually require multiple wires connected to different parts of human body and the participant usually cannot move freely during the capture. More recently such issues have been addressed by developing wearable sensors that are wireless and miniaturized. For example, Microsoft Band is a wrist worn sensor that can detect both heart rate, GSR and skin temperature in real time (See Figure 2.9). However, as suggested by Gunes and Schuller (2013) obtaining accurate measurement from biomedical sensors is still affected by human physical activities such as walking and running.

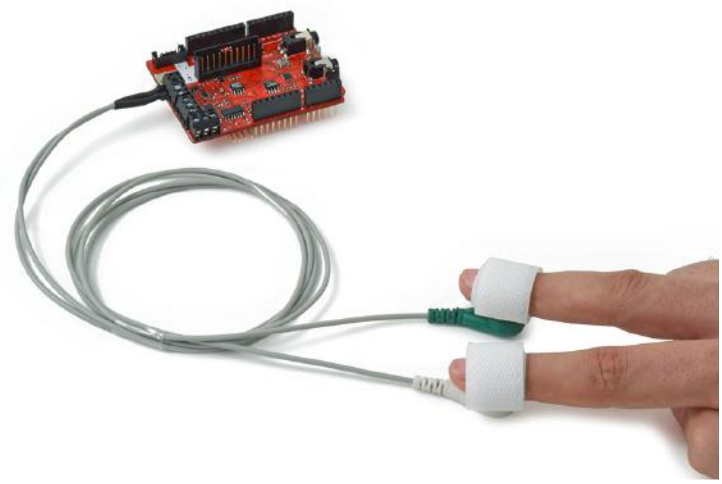


Figure 2.7: GSR Sensor (EHealth, 2013)

2.3.3 Annotating Affect Data

The process of annotating affect data is usually defined based on two criteria, i) the emotion model used to represent the data (Section 2.2) and ii) the method used to elicit emotions (Section 2.3.1). As discussed earlier, since the use of an appraisal model remains an open research question in automatic affect recognition, this section only focuses on how to use categorical and dimensional emotion models to annotate data.

For posed behaviours, a categorical model is commonly used. Usually there is no additional annotation step involved since posed behaviours are produced by the subject upon request. For induced and spontaneous behaviours, both categorical and dimensional emotion models are commonly used.

When using a categorical model approach, discrete emotion categories are used directly to annotate different emotional states. The six basic emotion categories developed by Ekman (1971) are the most widely used labels in automatic affect recognition research. In addition to the basic emotions, some datasets also include

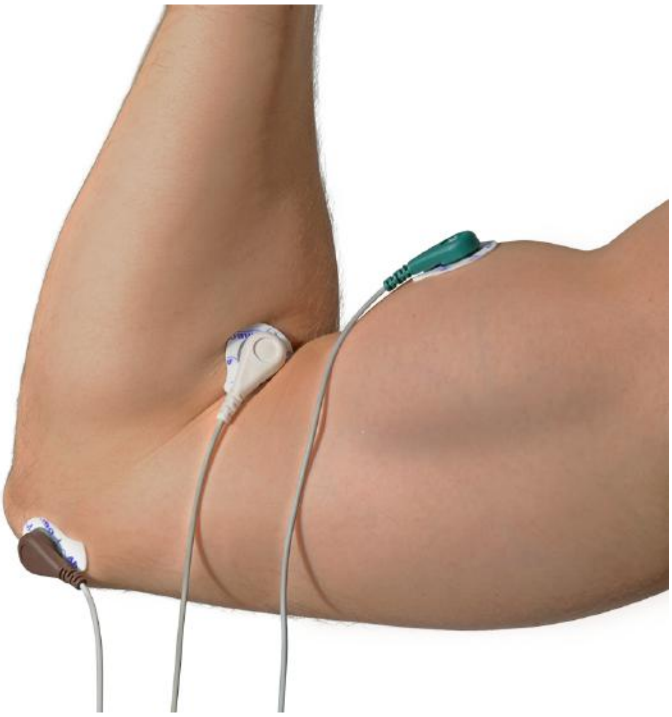


Figure 2.8: EMG Sensor (EHealth, 2013)

Figure 2.9: The Microsoft Band 2

labels to indicate people's cognitive states such as interest, puzzled, bored and frustration (Chen, 2000; Gunes and Piccardi, 2006).

When using a dimensional approach, both discrete and continuous labels can be used. For discrete labels, researchers have been using different intensity levels: either a set of words (e.g., negative, positive and neutral), or a ten-point Likert scale (e.g., 0-9 where 0 means a low value on the selected dimension), or an arbitrary range (e.g., integer value between -50 and +50). On the other hand, continuous labels usually use a real number between a predefined range (e.g., -1 to 1) to indicate various affective states along different dimensions (McKeown *et al.*, 2012; Ringeval *et al.*, 2013).

There are two possible approaches when it comes to annotating the data, subjective assessment (self assessment) and objective assessment. These methods can be used independently or used together. Subjective assessment asks participants to rate their own response to different stimuli through a recall process, while objective assessment requires external observers (human raters or annotators) to estimate the emotional state expressed by the subjects. In the affective computing field, recent research trends have been focused on recognising labels from objective assessment as shown in (Schuller *et al.*, 2011, 2012; Ringeval *et al.*, 2015b). One of the main challenges of objective assessment is to obtain high inter-observer agreement when multiple annotators are annotating the same data (Gunes and Schuller, 2013), especially when the continuous dimensional approach is adopted, since different annotators can annotate a different intensity for the same emotional state. To date, researchers have mostly chosen to take the average value among different annotations as the ground truth. Other methods, that take into account agreement and correlation measures have also been proposed. For instance, Nicolaou *et al.* (2011) measured the inter-observer correlation of each individual annotator and used it to calculate a weighted average as the final ground truth. Overall, as suggested by Gunes and Schuller (2013), obtaining reliable ground truth from both discrete and continuous dimensional annotations remains a challenging issue in affective computing.

Numerous tools with different functionalities have been developed to ease the annotation process. The **E**uropean distributed corpora project **L**inguistic **AN**notator (ELAN) (Brugman and Russel, 2004) is an annotation tool that allows users to create, edit, visualise and search annotations for video and audio data. It runs on all major operating system and is available in a number of different interface languages. In addition, it supports annotation on multiple levels such as word or sentence. The annotations can be created on multiple layers, called tiers, which can be hierarchically interconnected. ELAN was originally designed for the analysis of languages, sign languages, and gestures, but it had been used extensively in affect research (Valstar and Pantic, 2010).

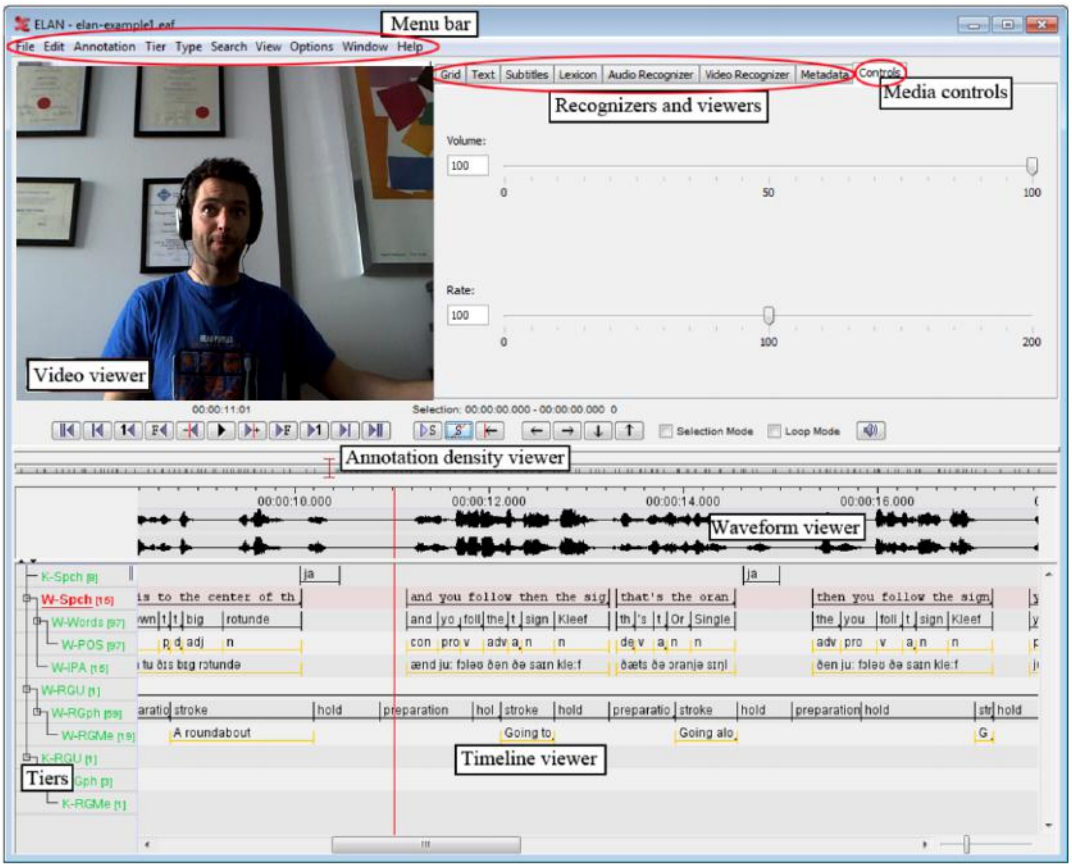
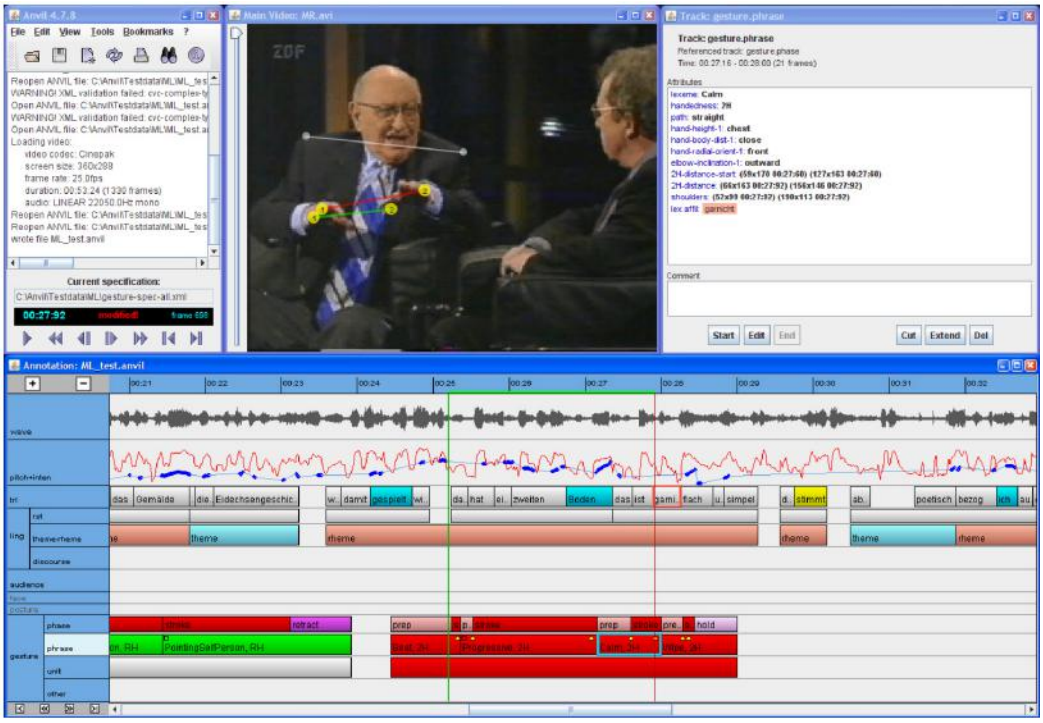


Figure 2.10: Screenshot of ELAN interface including the menu bar, the media player control, the tiers, and a number of viewers.

Another widely used video annotation tool is **AN**notation of **VI**deo and **L**anguage

(ANVIL) which was introduced by Kipp (2001). ANVIL is designed to facilitate annotation of audio-visual material. It supports multiple layers of annotation such as words, dialogue acts, postures shifts and gestures. It enables users to create their own coding scheme and review the colour-coded multi-layer annotation in a time-aligned fashion (See Figure 2.11). The tool is constantly updated. The latest version, ANVIL 5, now supports additional features such as 3D Motion capture player, subdivision track type, time point track type, histogram, Kappa coding agreement analysis, transition diagrams, association analysis and annotation management software. For more details the reader is referred to Kipp (2010).



```
Reopen ANVIL file: C:\Anvil\Testdata\ML\ML_tes
WARNING! XML validation failed: cvc-complex-t
Open ANVIL file: C:\Anvil\Testdata\ML\ML_test.a
WARNING! XML validation failed: cvc-complex-t
Open ANVIL file: C:\Anvil\Testdata\ML\ML_test.a
Loading video:
   video codec: Cinepak
   screen size: 360x288
   frame rate: 25.0fps
   duration: 00:53:24 (1330 frames)
   audio: LINEAR 22050.0Hz mono
Reopen ANVIL file: C:\Anvil\Testdata\ML\ML_tes
Reopen ANVIL file: C:\Anvil\Testdata\ML\ML_tes
wrote file ML_test.anvil
```

Current specification: C:\Anvil\Testdata\ML\gesture-spec-all.xml

00:27:92 modified! frame 698

Track: gesture.phrase
Referenced track: gesture.phase
Time: 00:27:16 - 00:28:00 (21 frames)

Attributes
lexeme: Calm
handedness: 2H
path: straight
hand-height-1: chest
hand-body-dist-1: close
hand-radial-orient-1: front
elbow-inclination-1: outward
2H-distance-start: (59x170 00:27:60) (127x163 00:27:60)
2H-distance: (66x163 00:27:92) (156x146 00:27:92)
shoulders: (57x99 00:27:92) (190x113 00:27:92)
lex affil: garnicht

Comment

Figure 2.11: ANVIL's graphical user interface.

The FEELtrace annotation tool was developed to enable annotators to track affective state via vocal and visual cues over continuous traces in dimensional space (Cowie *et al.*, 2000). The FEELtrace tool allows the annotator to watch the audio-visual recording and rate the perceived emotional state by moving the mouse pointer within the 2-dimensional valence-arousal space (See Figure 2.12). The value of the affective states has been confined to [-1, 1] where -1 represents very negative (valence) or very passive (arousal) and 1 represents very positive (valence) or very active (arousal). More recently, the General trace (Gtrace) annotation tool has been introduced to replace the FEELtrace tool with the ability to let people use their own dimensions and scales (Cowie and Sawey, 2011). The Gtrace tool lets the user annotate each dimension by moving the mouse cursor from left to right (See Figure 2.13).

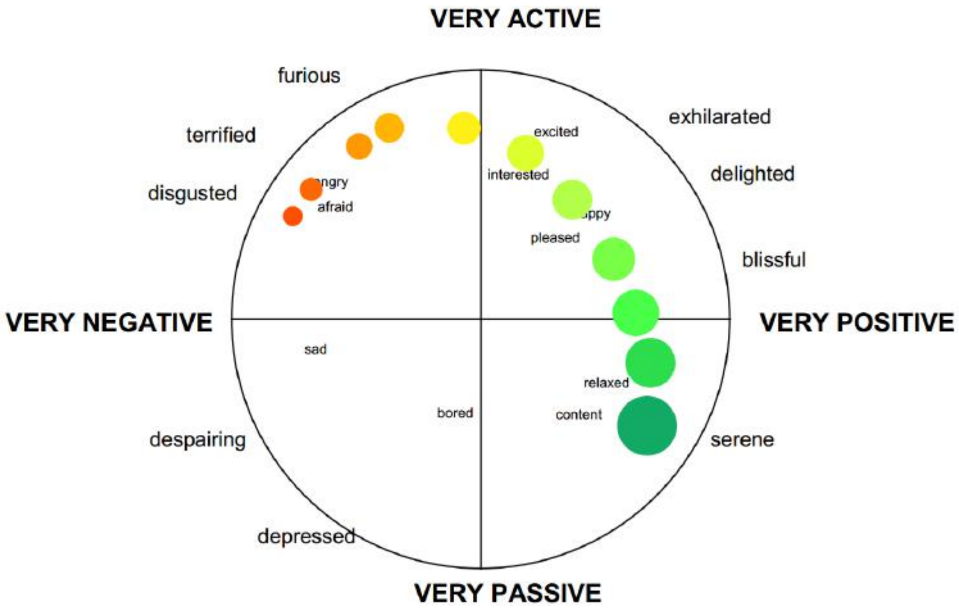


Figure 2.12: Feeltrace graphical user interface. Cursor colour changes from red/orange at the left hand end of the arc, to yellow beside the active/passive axis, to bright green on the negative/positive axis, to bluegreen at the right hand end of the arc (Cowie *et al.*, 2000).

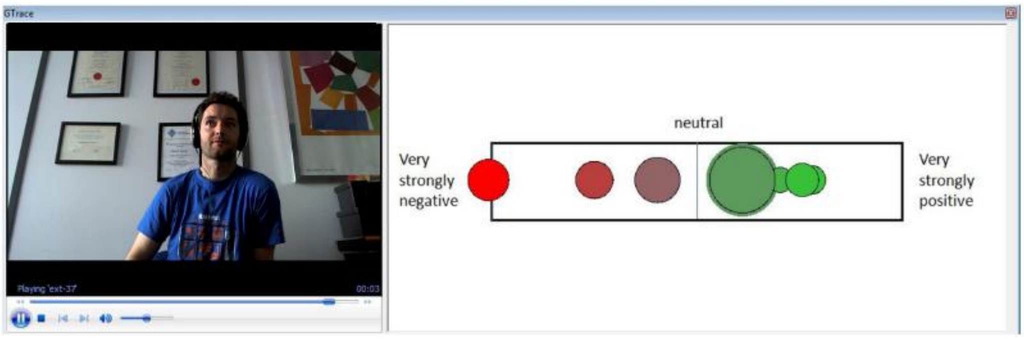
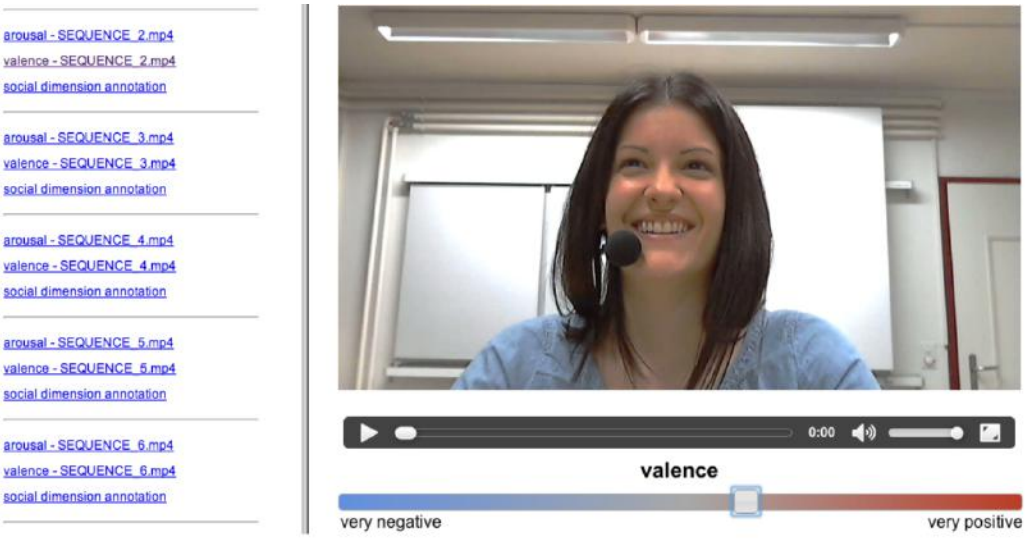


Figure 2.13: Gtrace graphical user interface (Cowie and Sawey, 2011)

A web-based annotation tool similar to Gtrace, called ANNEMO, (Figure. 2.14) has also been developed to enable remote annotation (Ringeval *et al.*, 2013). The arousal and valence dimensions are annotated using a slider with values from -1 to 1. An additional five dimensions (agreement, dominance, engagement, performance and support) can be annotated using a 7-Likert scale. The timestamps from the local machine are recorded when the slider value changes to avoid data transmission delays.



arousal - SEQUENCE_2.mp4
valence - SEQUENCE_2.mp4
social dimension annotation

arousal - SEQUENCE_3.mp4
valence - SEQUENCE_3.mp4
social dimension annotation

arousal - SEQUENCE_4.mp4
valence - SEQUENCE_4.mp4
social dimension annotation

arousal - SEQUENCE_5.mp4
valence - SEQUENCE_5.mp4
social dimension annotation

arousal - SEQUENCE_6.mp4
valence - SEQUENCE_6.mp4
social dimension annotation

Figure 2.14: ANNEMO: web-based annotation of affective and social behaviours (Ringeval *et al.*, 2013).

2.3.4 Existing Datasets

As discussed in Section 2.2, early research in affective computing field was mainly focused on using a categorical model with posed behaviours, so that the datasets captured in the early years only consist of acted affective behaviours. However, increasing evidence suggests that posed behaviours differ in *visual appearance*, *audio profile* and *duration of behaviour* compared to spontaneous behaviours (Zeng *et al.*, 2009). For instance, Whissell (1989) suggests that the choice of words and timing used in spoken language differ in posed and spontaneous behaviours. Cohn and Schmidt (2004) and Valstar *et al.* (2007) find that different types of spontaneous smiles exhibit smaller amplitude and longer duration compared to posed ones. These findings motivated several efforts toward capturing datasets consisting of spontaneous affective states that could be used for training and testing automatic affect recognition systems. Table 2.1 lists some representative and publicly available affect datasets that are reported in the literature. Since this thesis focuses on recognising affective state from the visual modality, only datasets with visual information are reviewed. For emotional speech datasets the reader is referred to Zeng *et al.* (2009), El Ayadi *et al.* (2011) and Koolagudi and Rao (2012) for more details.

For datasets concerning posed affective behaviours, the following datasets should be noted. The Cohn-Kanade (CK) dataset (Kanade *et al.*, 2000) is the most widely used dataset for facial expression recognition. It includes 97 adults across 3 races. In total, 486 video sequences are captured where each sequence consists of a different number of frames (from 9 to 60 frames). The dataset was extended by (Lucey *et al.*, 2010) to include an additional 26 subjects and 107 sequences. All sequences are fully FACS coded and labelled with categorical emotion labels (angry, contempt, disgust, fear, happy, sadness and surprise). The MMI dataset (Pantic *et al.*, 2005) is one of the first on-line accessible and searchable facial expression datasets. Originally, it only contained posed facial expressions, but was expanded by Valstar and Pantic (2010) to also include induced facial behaviours. The original MMI dataset consists of 19 subjects across 3 raters. In total, 600 frontal and 140 dual-view static images are captured along with 30 profile-view and 750 dual-view video sequences. The 'dual-view' capture refers to the capture of frontal and profile views at the same time. The extended MMI dataset adds an additional 25 subjects, where each subject was recorded for 5 minutes. All data in the MMI dataset is FACS coded. The six basic emotions are used to label the posed facial expressions, while only a subset (happiness, disgust and surprise) is used to label the induced ones. The FABO dataset developed by Gunes and Piccardi (2006) is one of the first datasets that contains videos of both facial expressions and body gestures. In total, 23 subjects were recorded, where each subject was asked to perform a set of pre-defined facial

Table 2.1: Comparison of datasets for human affect recognition. 2D: Colour information. 3D: Colour and depth information. A: Audio information. B: Biomedical information. C:Continuous annotation. D: Discrete annotation

dataset	**Elicitation Method**	**Modality**	**Emotion Model**	**Labelling**	**Environment**
Cohn-Kanade (Kanade *et al.*, 2000)	Posed	2D	Categorical	N/A	Controlled
MMI (Pantic *et al.*, 2005)	Posed Induced	2D	Categorical	Objective	Controlled
UT Dallas (O'Toole *et al.*, 2005)	Induced	2D	Categorical	Objective	Controlled
BU-3DFE (Yin *et al.*, 2006)	Posed	3D	Categorical	N/A	Controlled
FABO (Gunes and Piccardi, 2006)	Posed	2D	Categorical	N/A	Controlled
SAL (Douglas-Cowie *et al.*, 2007)	Induced	2D/A	Dimensional (C)	Objective	Controlled
BU-4DFE (Yin *et al.*, 2008)	Posed	3D	Categorical	N/A	Controlled
Bosphorus (Savran *et al.*, 2008)	Posed	3D	Categorical	N/A	Controlled
Vera am Mittag (Grimm *et al.*, 2008)	Spontaneous	2D/A	Categorical Dimensional (C)	Subjective	Controlled
BIWI 3D (Fanelli *et al.*, 2010)	Posed	3D/A	Categorical	Subjective	Controlled
Cam3D (Mahmoud *et al.* (2011)	Induced	3D/A	Categorical	Objective	Controlled
ICT-3DRFE (Stratou *et al.*, 2012)	Posed	3D	Categorical	N/A	Controlled
SEMAINE (McKeown *et al.*, 2012)	Induced	2D/A	Categorical Dimensional (C)	Objective	Controlled
MAHNOB-HCI (Soleymani *et al.*, 2012)	Induced	2D/A/B	Categorical Dimensional (D)	Subjective	Controlled
DEAP (Koelstra *et al.*, 2012)	Induced	2D/B	Dimensional (D)	Subjective	Controlled
BP4D-Spontaneous (Zhang *et al.*, 2013)	Induced	3D	Categorical	Subjective	Controlled
AViD (Valstar *et al.* (2013)	Induced	2D/A	Dimensional (C)	Objective	Uncontrolled
RECOLA (Ringeval *et al.*, 2013)	Spontaneous	2D/A/B	Categorical Dimensional (C)	Subjective Objective	Controlled

expressions and body gestures. The labels include six basic emotions and four non-basic affective states (uncertainty, anxiety, boredom, and neutral). The BU-3DFE dataset (Yin *et al.*, 2006) is one of the first publicly available facial expression datasets that contains both colour and depth information. The dataset was captured using a 3D scanner and was labelled using seven basic categorical (six basic emotions and neutral) emotions. Each expression includes four different intensity levels (low, middle, high and highest) except for the neutral expression. The dataset includes 100 subjects and a total of 2,500 3D facial expression models. However, due to the speed limitation of the 3D scanner used during the capture, the BU-3DFE only consists of static 3D information. This issue was addressed by the development of BU-4DFE dataset (Yin *et al.*, 2008) which contains 3D dynamic facial sequences. The dataset consists of 101 subjects where each subject is asked to perform six basic expressions, resulting in a total of 606 3D dynamic sequences. Different from aforementioned 3D datasets, both the Bosphours and the ICT-3DRFE datasets capture AU related expressions (see Figure 2.6 for some examples) in addition to the basic emotions. In particular, the Bosphorus 3D face dataset (Savran *et al.*, 2008) consists of 105 subjects performing 34 facial expressions under various head poses (13 yaw, pitch and cross rotations) and different face occlusions (hand, hair and eyeglasses), while the ICT-3DEFE dataset (Stratou *et al.*, 2012) includes 23 subjects and 15 expressions captured under different illumination conditions. The BIWI 3D (Fanelli *et al.*, 2010) dataset is the first dataset that contains both audio and 3D facial information captured in an affective communication setting. During the capture, each participant is asked to first read a sentence from text in neutral expression, and then repeat the same sentence after watching a clip extracted from a feature film where the sentence is acted by professional actors. This results in 1109 sentences spoken by 14 native English speakers.

Considering induced affective behaviours, the UT Dallas dataset (O'Toole *et al.*, 2005) is one of the first publicly available datasets that features induced emotions.

Different emotional states are elicited by asking participates to watch different video clips. The data is then labelled using 10 discrete emotion categories (6 basic emotions, puzzlement, laughter, boredom and disbelief). The SAL (Douglas-Cowie *et al.*, 2007) dataset is the first multi-modal dataset that was annotated continuously using the dimensional approach (arousal and valence). The SAL data consists of recordings of human-computer conversations elicited through a *Sensitive Artificial Listener* (SAL) interface. The interface is build around four personalities (happy, gloomy, angry and pragmatic) where each personality is represented by an avatar. The idea is that each avatar draws the participant into their own emotional state through a set of predefined responses. In total, 4 subjects and over 4 hours of data was captured. The SEMAINE dataset (McKeown *et al.*, 2012) also uses the same SAL induction technique, but differs in recording quality, size and annotation information. It includes 20 participants and over 6 hours of data was captured. It is annotated continuously on five dimensions including valence, arousal, dominance, expectation and intensity, where the intensity is used to indicate "*how far the person is from a state of pure, cool rationality*". In addition it also includes labels on basic emotions (i.e. fear, anger and happiness), epistemic states (i.e. certain/not certain, agreeing/not agreeing and interest/not interest), interaction process analysis (i.e. shows solidarity, shows antagonism and shows tension) and validity (i.e. breakdown of engagement and anomalous simulation). It has been used for the first and second Audio/Visual Emotion Challenge (AVEC 2011, 2012) (Schuller *et al.*, 2011, 2012).

The Cam3D dataset Mahmoud *et al.* (2011) is the first induced dataset that contains both 3D and audio information. The data is induced through human-computer and human-human interactions. In total 16 participants were captured and 108 segments of natural complex mental states were extracted. The segments were labelled in terms of agreeing, bored, disagreeing, disgusted, excited, happy, interested, neutral, sad, surprise, thinking and unsure. The MAHNOB-HCI (Soleymani *et al.*, 2012) and DEAP (Koelstra *et al.*, 2012) datasets are two of the first datasets that contain biomedical information and are annotated using the dimensional approach. The AViD (**A**udio-**Vi**deo **D**epressive) dataset (Valstar *et al.*, 2013) is the first dataset captured in an uncontrolled environment which includes 340 video clips of participants performing a HCI task. The data is annotated continuously in terms of arousal, valence and dominance. In addition, the depression level is also labelled with a single value per recording using a self-assessed depression questionnaire. It was chosen for the third and fourth AVEC (AVEC 2013, 2014) (Valstar *et al.*, 2013, 2014). For spontaneous affective state, two of the most widely used datasets are the Vera am Mittag dataset and the RECOLA dataset. The Vera am Mittage dataset (Grimm *et al.*, 2008) consists of 12 hours of audio-visual recordings of the German TV talk show 'Vera am Mittag'. The dataset is segmented into broadcasts, dialogue acts and utterances. It is annotated continuously on valence, arousal and dominance dimensions as well as using six basic emotions. The RECOLA dataset is the first spontaneous dataset that captures both audio-visual and biomedical data. The dataset is designed to capture spontaneous interactions from a remotely performed collaborative task. In total, 3 hours and 50 minutes of data and 46 participants were captured. Both subjective and objective assessment are used to annotate the data. The objective assessment is used to continuously annotate the data along the valence and arousal dimensions, while subjective assessment is used to indicate a participant's own emotional state at different stages of the capture. It was used for the fifth AVEC (AVEC 2015) (Ringeval *et al.*, 2015b). For clarity, the subset of the SEMAINE dataset used in the AVEC 2012 challenge is referred as the AVEC 2012 dataset and the subset of the RECOLA dataset used in the AVEC 2015 challenge is referred as the AVEC 2015 dataset. Both AVEC datasets are divided into training, development and testing partitions.

2.3.5 Discussion

Various affect datasets have been captured to fullfill the needs of training and testing automatic affect recognition systems. These datasets differ in how the affective state is elicited, what emotion model is used to represent the affective state and what modalities are being captured. As suggested by Gunes and Schuller (2013) the new emerging trend in continuous affect data acquisition is to focus on multi-modal and multi-speaker interactions rather then human-computer interactions. In addition, with the recent availability of affordable depth sensors, there has been a growing interest in collecting multi-modal affect datasets that contain 3D information. However, to the best of our knowledge, there still does not exist any dataset that includes recording of spontaneous behaviours with 3D information that is also annotated continuously using the dimensional approach. The lack of such a dataset makes it impossible to study how 3D information could be used for spontaneous affect recognition. To address these issues, a multi-modal multi-speaker 3D spontaneous affect dataset was captured as part of this thesis. More details on how this dataset was created can be found in Chapter 3. In addition, since this thesis focuses on continuous affect detection, two of the most widely used datasets, AVEC 2012 and 2015 dataset are also chosen for the experiments carried out in this thesis.

2.4 Affect Recognition

Recognising affective state from different input signals raises a number of research challenges. These include feature extraction, machine learning and performance evaluation (See Figure 2.15). In the following sections, each of the aforementioned challenges are reviewed. Since this research mainly focuses on using the visual modality to recognise affective states, only visual features will be reviewed. For features concerning other modalities such as audio and biomedical, the reader is referred to Zeng *et al.* (2009), Gunes (2010) and Weninger *et al.* (2013) for more details.

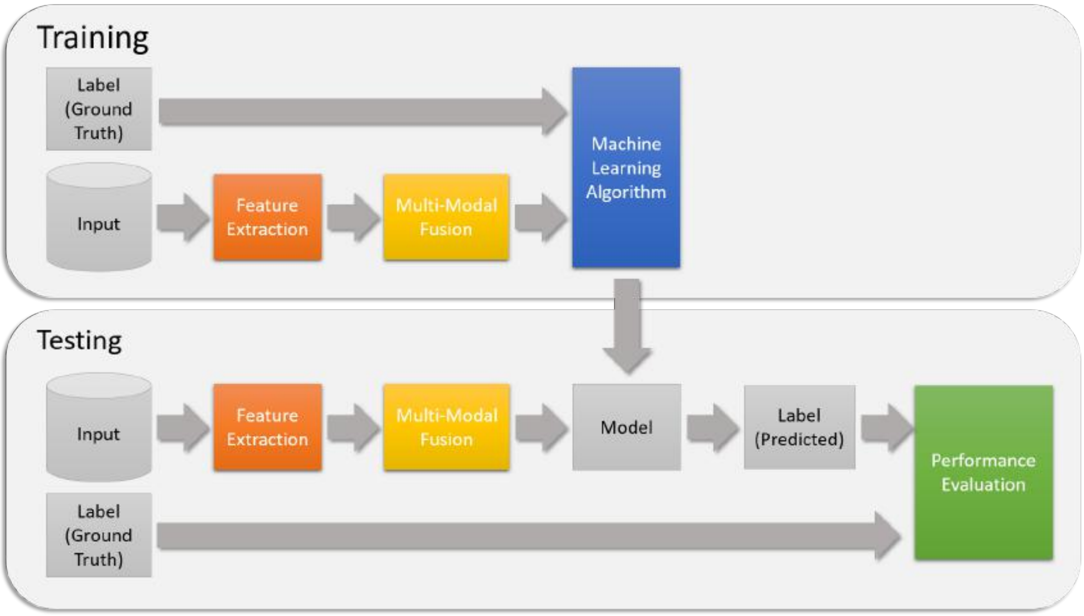


Figure 2.15: Overview of A Typical Affect Recognition System

2.4.1 Feature Extraction

Feature extraction is the process of transforming the input data into a lower dimensional space while remaining its representative characteristics. Discriminative features are usually desired to achieve good recognition results. This means the features should have high intra-class variation (features should be different for different classes) and low inter-class variation (features should be similar for same class). In this section, features extracted from both 2D and 3D visual signals are discussed. In addition since the face is the most visible part of the human body that reveals emotions (Ekman and Rosenberg, 1997) and people constantly read others' facial expression in order to understand how others feel (Schmidt and Cohn, 2001) only facial features are reviewed.

In order to extract features from a face, the location of the face must be first detected, a process known as face detection. Numerous face detection algorithms have been proposed for 2D images, and comprehensive details of state-of-the-art in face detection can be found in Zhang and Zhang (2010) and Zafeiriou *et al.* (2015). Among different face detection algorithms, the work carried out by Viola and Jones (2004) is one of the most popular ones due to its simple, fast and open source implementation. Several face detection algorithms have also been proposed that use depth images. For instance, the aforementioned classic Viola-Jones face detection algorithm has been extended to depth by Burgin *et al.* (2011). Mattheij *et al.* (2012) employs Haar-like region features on the integral image representation of depth images for robust and accurate face detection. Li *et al.* (2013) proposed another face detection algorithm specifically designed for low resolution 3D sensors by using the Iterative Closest Point method (ICP) to estimate the location of the tip of the nose and thus localise the face. After the face is located, the next step is to extract features from the visual signals.

2.4.1.1 2D Visual Facial Features

In general, facial features from 2D visual signals can be divided into two main categories: geometric features and appearance features.

Geometric features aim to incorporate knowledge from cognitive science to analysis facial variations under different affective states. Facial shape and activity can be represented by a set of points called landmarks as shown in Figure 2.16. The problem of localising these landmarks is called Face Alignment. For a comprehensive review of this field the reader is referred to Wang *et al.* (2014) and Jin and Tan (2016). Geometric features are usually derived from these facial landmarks. The most frequently used geometric feature representation is to simply concatenate the x and y coordinates of a number of landmark points. In order to reduce the head pose variation and identity bias, normalisation is usually applied to the landmarks. This can be achieved by removing the similarity parameters of the shape model and by using the landmarks from a neutral face image (Lucey *et al.*, 2007, 2009). In addition to the landmark coordinates, the distance and angle between certain landmarks and parameters from the shape model are also used. For example, Huang *et al.* (2010) developed a triangular-based facial features descriptor to calculate the distance between certain facial landmarks. Nicolle *et al.* (2012) adopts the shape model parameters directly as geometric features. Valstar and Pantic (2012) use the landmark coordinates along with the length and angle of all pairwise points in space, and the difference between these features with respect to their value in a neutral face. Similarly, Ringeval *et al.* (2015b) proposed to use a combination of different sets of geometric features based on facial landmarks. The first sets include the difference between the currently aligned landmarks and those from the mean shape, and also between the aligned landmarks in the previous frame. The second set computes the Euclidean distance and angle between pairs of points in three different regions. The third set calculates the distance between the median of stable landmarks and each aligned landmark in the current frame. A number of studies have shown that the geometric features are extremely useful at predicting the valence dimension when compared to appearance and audio features (Ringeval *et al.*, 2015b). Geometric features are robust to lighting conditions since the focus is on coordinates of landmarks rather than the intensity of the pixel. However, they are very sensitive to facial landmarks registration errors as they are calculated purely based on the landmark coordinates. In addition, although the geometric features could describe temporal variations, they may not be able to capture subtle expressions using a limited number of facial points.

Appearance features aim to measure the motion and change in texture for affect recognition. There are two main categories for appearance features: filter bank based features and histogram-based features. Two of the most widely used filter-bank-based features are Gabor filters and Haar-like filters.

The Gabor representation is obtained by convolving the input image with a set of Gabor filters of different frequencies and orientations. Typically in the literature this corresponds to 8 orientations, and a number of frequencies from 3 to 9. It has been

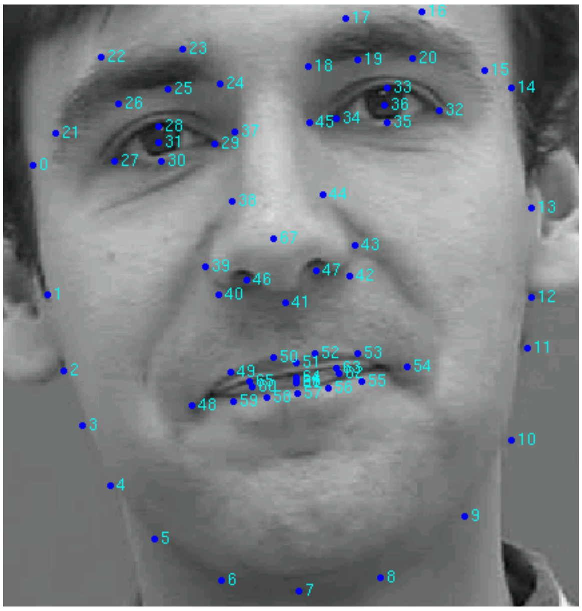


Figure 2.16: Sample Landmarks (TalkingFaceVideo, 2002)

shown that the frequency and orientation representation of Gabor filters are similar to the response of simple cortical cells (Marčelja, 1980; Daugman, 1985). A Gabor filter with a given orientation results in a strong response for a specific location in the target image that exhibits edges and texture changes in the given direction. This means that when the filter frequency and direction match the image structure, they can be sensitive to finer wave-like image structures such as those in facial expressions. When used in affect recognition, only Gabor magnitudes are commonly used, as they have been proven to be robust to face misalignment (Stewart *et al.*, 2006; Gritti *et al.*, 2008; Mahoor *et al.*, 2011). However, the Gabor representation is computationally costly due to convolution with a large number of filters (e.g., 8 orientations and 3 frequencies implies 24 filters) and the dimensionality of the convolution output is high.

The Haar-like representation (Papageorgiou *et al.*, 1998) considers adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculates the difference between these sums. The Haar-like filter responds to coarser image features is robust to shift, scale and rotation variation, and is very fast to compute. However, it is not responsive to finer texture details and as a result it is limited in the ability to detect expressions with more obvious facial muscle actions.

The general procedure for extracting histogram-based representations consists of three stages: first local features are extracted and encoded in a transformed image, then the transformed image is divided into uniform regions, and finally the local histogram is generated by pooling the features in each region. The final feature vector is formed by concatenating all local histograms together. Some of the most commonly used histogram-based representations in affect recognition are Local Binary Pattern (LBP), Local Phase Quantisation (LPQ), Local Gabor Binary Patterns (LGBP), Local Binary Patterns on Three Orthogonal Planes (LBP-TOP), Local Phase Quantisation on Three Orthogonal Planes (LPQ-TOP) and Local Gabor Binary Patterns on Three Orthogonal Planes (LGBP-TOP).

The local binary pattern of a pixel is defined as an 8-bit binary number that can be obtained by comparing each pixel intensity against the intensity of its neighbouring pixels. The original LBP feature (Ojala *et al.*, 1994) is represented as a histogram where each bin corresponds to the number of one of the different possible local binary patterns, resulting in a 256-dimensional feature vector. Various extensions have been proposed for the original LBP feature, and one of the most commonly used extensions is called *uniform pattern LBP* (Ojala *et al.*, 2002). The extension was inspired by the fact that some binary patterns occur more commonly in texture images then others. A uniform pattern is defined as the binary pattern that consists of at most two bitwise transitions from 0 to 1. Eliminating binary patterns that do not meet the uniformity condition reduces the dimension of the original LBP feature from 256 to 59. The main advantages of LBP features are their robustness to lighting conditions and shifts while maintaining computational simplicity (Shan *et al.*, 2009). However, they are less robust to rotation, and as a result a normalisation that rotates the face to upright position is usually required

(Schuller *et al.*, 2011, 2012).

The LPQ feature was originally proposed as a descriptor for texture classification (Ojansivu and Heikkilä, 2008). The LPQ feature uses phase information computed locally over a predefined rectangle for every pixel using the short-term Fourier transform (STFT). The local Fourier coefficients are computed at four frequencies and the phase information in the Fourier coefficients for each frequency is quantised by keeping the signs of the real and imaginary parts of each component, resulting in a 8-bit binary representation. A histogram is then constructed similar to the LBP features where each bin corresponds to one of the specific binary patterns. This forms a 256-dimensional feature vector. One advantage of LPQ feature is its robustness to image blurring produced by a point spread function (Mandal *et al.*, 2015, p. 149). The LPQ feature has been used for both facial action detection (Jiang *et al.*, 2011) and affect recognition (Valstar *et al.*, 2013)

The LGBP feature was first proposed by Senechal *et al.* (2012). The computation of LGBP features is similar to the LBP feature, with the difference that the LGBP feature approach is to first apply a set of Gabor filters (typically with 3 frequencies and 6 orientations) to the input image before the local binary pattern is computed. The LGBP feature exhibits the advantage over Gabor representations of being robust to illumination changes and misalignments. However, its use in affect recognition is less common compared to LBP and LPQ features due to its high computation cost and high dimensionality. For instance, the typical configuration results in 18 Gabor images, and after concatenating the LBP histogram from each Gabor images, the final LGBP feature dimension is 18 times bigger than the LBP representation.

The histogram-based representations discussed above are usually robust to spatial illumination variations to a degree and invariant to global illumination, as they are extracted from small patches. However, one disadvantage is that the aforementioned features are all static features since they are all calculated based on a single image, while facial expression is a dynamic event. For instance, someone with a particular physiognomy could look like he/she is smiling when in fact there was no facial action at all (Mandal *et al.*, 2015, p. 149). To overcome this problem, a dynamic extension of the LBP feature was proposed by Zhao and Pietikainen (2007). To make the calculation computationally efficient, the LBP features were only computed on Three Orthogonal Planes (TOP): XY, XT and YT, resulting in the so called LBP-TOP feature. The basic idea behind TOP is that a video sequence is usually viewed as a stack of XY planes along the T axis, but can also be treated as a stack of XT planes in axis Y and YT planes in the X axis, respectively. The XT and YT planes contain information about the space-time transitions of the textures. The LBP feature is calculated on each of the three planes and the final feature vector is formed by concatenating the histogram from each of the planes. The same extension was later proposed for LPQ features (Jiang *et al.*, 2011) and LGBP features (Almaev and Valstar, 2013). There are three main drawbacks of the TOP features. Firstly, the dimensionality of the TOP features are usually much larger than their static counterparts. Secondly, the computation time is much longer for TOP features, especially for LPQ-TOP and LGBP-TOP features. Finally, since the TOP features are computed over a fixed temporal window, the same facial expression produced at different speeds could result in different feature representations, thus increasing intra-class variability (Mandal *et al.*, 2015, p. 150).

A number of studies on facial action unit detection have shown that the performance of LBP, LPQ and LGBP features have been significantly improved after their TOP variants were used (Jiang *et al.*, 2011; Almaev and Valstar, 2013). In addition, the TOP variants are proven to be more robust to rotational misalignments when compared to their static counterparts (Almaev and Valstar, 2013). When it comes to continuous affect recognition, researchers have applied different histogram representations to different datasets, but there is no clear indication which representations achieved the best result.

2.4.1.2 3D Visual Features

According to Sandbach *et al.* (2012), features used in 3D facial expression recognition systems can usually be divided into distance features, patch features, model features, and 2D representation features. This section briefly reviews each type of feature. For a more detailed discussion on these features, the reader is referred to Sandbach *et al.* (2012).

The distance features are similar to the geometric features used for 2D images which are usually calculated based on facial landmarks. For instance, Soyel and Demirel (2007) use the distance vectors derived from the 3D distribution of facial landmarks provided by the BU-3DFE dataset to classify facial expressions. Tang and Huang (2008) calculate the distance and slope between certain pairs of facial landmarks and use this as a feature set to classify six basic emotions using the BU-3DFE dataset. Similarly, the work carried out by Li *et al.* (2010) uses distances, angle and slope related to the movement of a specific facial part and the shape of the eyes and mouth to recognise discrete emotions for the BU-3DFE dataset.

The patch features are used to capture the shape of the face over a small region around either every point in a mesh or around facial landmarks. Wang *et al.* (2006) proposed to fit a smooth polynomial patch at each point in the mesh and use the parameters derived from these patches to differentiate six basic emotions on a custom dataset. Alternatively, Maalej *et al.* (2010) uses the shape information of the closed patch found around each facial landmark to recognise the six basic emotions using the BU-3DFE dataset.

The aim of model-based features is to fit a 3D face model using the depth information, and to use the parameters derived from the 3D face model to recognise different emotions. For instance, the work carried out by Ramanathan *et al.* (2006) uses a Morphable Expression Model (MEM) to recognise four expressions: neutral, happy, sad and angry on a custom dataset. The depth data is fitted by minimising the energy function between certain triangular meshes. The morphing parameters produced during this process are used as the features. Gong *et al.* (2009) proposed to use the Basic Facial Shape Component (BFSC) to recognise emotions in the BU-3DFE dataset. The depth data was first aligned using Iterative Closest Point (ICP), then the BFSC was fitted to each mesh, and finally the difference between the aligned mesh and the BFSC was used to form the feature vectors.

The 2D representation-based features usually can be divided into two categories based on how the depth data is converted to 2D representations: direct conversion and indirect conversion. The direct conversion uses the z value (distance value) directly at each x, y position to form a depth map, while the indirect conversion applies a transformation to depth data to form a 2D representation. For instance, Berretti *et al.* (2011) carried out an experiment on BU-3DFE dataset using the direct conversion approach. After conversion, the Scale-Invariant Feature Transform (SIFT) algorithm was applied at each of the automatic detected landmarks to extract local features. Vretos *et al.* (2011) calculates Zernike moments on the histogram equalised depth map and used these as the features to classify basic emotions on both BU-3DFE and Bosphrous datasets. The work by Zhen *et al.* (2013) explored the use of LBP-TOP and LPQ-TOP on a depth map to classify the six basic emotions for the BU-4DFE dataset. On the other hand, Rosato *et al.* (2008) investigated the use of conformal mapping to convert 3D meshes to 2D planar meshes. The 2D planar meshes are then used to generate a set of labels to describe the small surface variation during the motion of facial surfaces, and finally the distributions of different labels are used as features to classify different emotions using the BU-3DFE dataset. Savran *et al.* (2012b) proposed to convert the depth data into a 2D representation using differential geometry-based features and experiments were carried out on the Bosphorus dataset for the AU detection task.

2.4.2 Machine Learning for Affect Recognition

After an appropriate feature representation has been extracted, it is the task of the machine learning component to learn how to match the feature representations to the target labels. In this section we introduce the machine learning techniques used later in the thesis. Specifically, the Support Vector Regression (SVR) is used to investigate the performance of various histogram-based visual features, the Convolutional Neural Network (CNN) is used to study the performance of automatic generated visual features, and the Long-Short Term Memory (LSTM) neural network is used in the multi-modal fusion.

To formalise the machine learning problem, assume there exists a hidden function $f : X \to Y$ that for a input instance $x \in X$, generates an output instance $y \in Y$. Machine learning techniques try to learn a hypothesis function $h : X \to Y$ as close as possible to the hidden function f. Depending on the learning tasks, machine learning techniques for affect recognition are usually divided into two categories: *classification* and *regression*. If Y are discrete values then this is defined as a classification problem. When there are only two discrete labels, this is often called *two-class classification*. When there are more discrete labels, this is often referred to as *multi-class classification*. Classification machine learning techniques are commonly used for predicting categorical and discrete dimensional affective states. If Y are continuous values, then this is defined as a regression problem. Regression machine learning techniques are commonly used for continuous dimensional affect recognition. The most common way to learn the hypothesis function h is achieved by defining a cost function that will assign a value to indicate the difference between ground truth labels and labels being predicted by the hypothesis function h. Depending on the type of cost function, the goal of machine learning algorithms is to either minimise or maximise the cost function. Different machine learning techniques have been successfully applied to affect recognition such as Support Vector Machine

(SVM), K-Nearest Neighbour Classifiers (KNN), Decision Tress, Conditional Random Fields (CRF), Hidden Markov Models (HMM), Convolutional Neural Networks (CNN) and Long Short Term Memory Recurrent Neural Networks (LSTM). The following section briefly introduces two of the most commonly used machine learning techniques for affect recognition.

2.4.2.1 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a popular machine learning technique that has been used widely for both classification and regression problems. It was first introduced by Vapnik (1995) based on the Vapnik-Chervonenkis' (VC) theory (Vapnik and Chervonenkis, 1968, 1971). The goal of an SVM is to find the optimal hyperplane which maximizes the margin of the training data. This is shown in Figure 2.17 where the positive sign and negative sign indicates two different classes. In order to separate the data, a line could be used, commonly called a *hyperplane*. In this case, three possible hyperplanes are highlighted, shown in blue, orange and green. The hyperplane is usually defined as $h(x) = w^T x + b$ where w is called the *weight vector* and b is called the *bias*. The *margin* is defined as the distance between the closest training examples and the hyperplane. By varying the value of w and b, one can find a set of parameters that maximizes the margin. In this case, the blue hyperplane has the largest margin and is chosen as the optimal hyperplane. The closest training examples (A, B, C) on the dash line are called *Support Vectors*. To find the best set of w and b, the Lagrange multipliers can be used to construct the Lagrange functions (Cortes and Vapnik, 1995), and then the Sequential Minimal Optimization (SMO) algorithm developed by Platt *et al.* (1998) can be used to solve the Lagrange functions. When used for classification, the cost function of an SVM is defined as:

$$\ell(\hat{y}, y) = max(0, 1 - \hat{y} \cdot y)$$

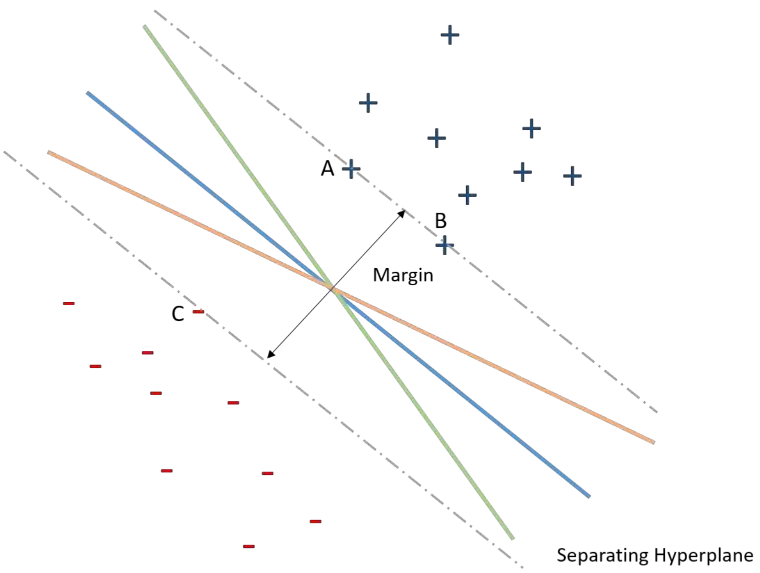


Figure 2.17: Data points and hyperplanes (Cortes and Vapnik, 1995)

where y is the ground truth label and $\hat{y}$ is the predicted label. This cost function is referred to as the *hinge loss* function. The hinge loss is 0 when y and $\hat{y}$ have the same sign (meaning the classifier's prediction is the same as the ground truth) and increases linearly with y when they have different signs. When used for the regression task, the ϵ-*insensitive loss* function proposed by Smola and Vapnik (1997) is commonly used. It is defined as:

$$\ell(\hat{y}, y) = \begin{cases} 0 & \text{if } |y - \hat{y}| \leq \epsilon \\ |y - \hat{y}| - \epsilon & \text{otherwise} \end{cases}$$

where ϵ is used to define the number of errors the loss function ignores. By minimising the loss function, a hyperplane can be found that best fits the given training data as shown in Figure 2.18. When the training data that is not linearly separable, a technique called the *Kernel Trick* is employed. The Kernel Trick works by projecting the training data into a high-dimensional space where the data can be linearly separated using kernel functions. Some of the most commonly used kernel functions including Polynomial Function, Radial Basis Function (RBF) and Sigmoid

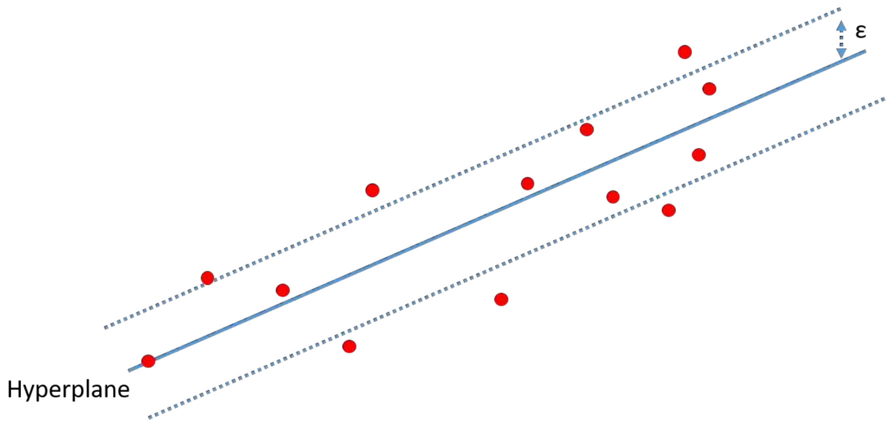


Figure 2.18: Support Vector Machine for Regression

Function.

2.4.2.2 Neural Network

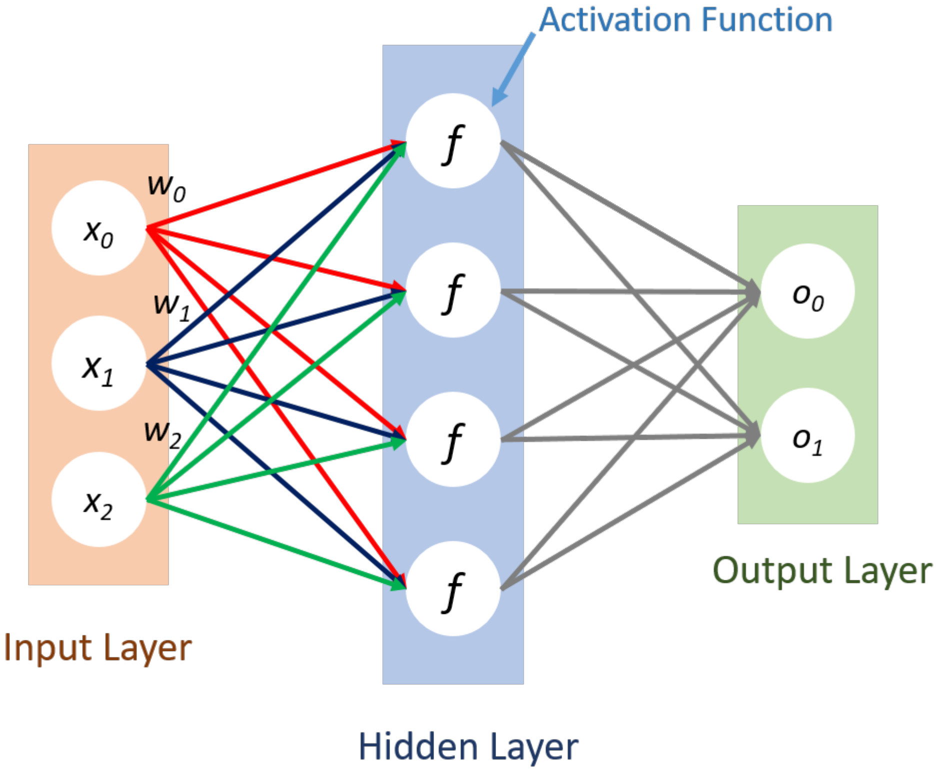
In recent years, neural networks, especially deep neural networks (also known as deep learning) have attracted a great amount of research interest. They have been applied to solve different research problems such as object classification, speech recognition and machine translation, and have achieved state-of-the-art performance. A Neural Network is inspired by the structure of the biological brain, specifically it uses a large collection of neural units to model the way a brain solves problems. Neural Networks are typically organised in layers. These generally consists of three types of layers: *input layer*, *hidden layer* and *output layer*. If there is more then one hidden layer, this is usually referred to as a *deep neural network*. The hidden layer usually consists of a number of interconnected neurons (nodes) which contain an *activation function*. The output of the activation function is calculated as $f(\sum w_i x_i + b)$ where x is the input vector, w is the weight vector and b is the bias. This is shown in Figure 2.19. Some of the most commonly used activation functions are shown in Table 2.2.

The goal of the neural network is to learn the weight vectors through training data that minimise the error between predictions and ground truth. This is achieved

Table 2.2: Commonly used activation functions and their corresponding derivatives

Activation Function Type	**Function**	**Derivative**
Binary threshold	$y(z) = \begin{cases} 1 & z \geq \theta \\ 0 & z < \theta \end{cases}$	$y' = 0$
Identity	$y(z) = z$	$y'(z) = 1$
Sigmoid	$y(z) = \frac{1}{1+e^{-z}}$	$y'(z) = y(z)(1 - y(z))$
Hyperbolic Tangent (tanh)	$y(z) = tanh(z)$	$y'(z) = 1 - tanh^2(z)$
Rectified Linear Unit (ReLu)	$y(z) = \begin{cases} z & z \geq \theta \\ 0 & z < \theta \end{cases}$	$y(z) = \begin{cases} 1 & z \geq \theta \\ 0 & z < \theta \end{cases}$

Figure 2.19: A 2-layer Neural Network



using the so called *backpropagation* algorithm. The algorithm works by first initialising the weight vector randomly, then computing the error between the output from forward propagation of the ground truth, and then calculating the gradient of weight layer-by-layer using the chain rule, which allows the error to propagate back from output to input. The weight vector is then updated by subtracting a portion of the gradient. The portion is usually refered as the learning rate. The algorithm iterates the process until the error no longer changes or a predefined error threshold has been reached. More details about the backpropagation algorithm can be found in the paper by Rumelhart *et al.*

The neural network structure shown in Figure 2.19 is called a *Multilayer Perceptron (MLP)* where each layer is fully connected to the next one except for the input layer. The *Convolutional Neural Network (CNN)* is another popular neural network structure that has attracted a great amount of research interest in recent years given its superior performance in object recognition tasks (Krizhevsky *et al.*, 2012). Two of the key concepts of a CNN are the convolutional layer and pooling layer (sub-sampling layer). Unlike the fully connected layer, in the convolutional layer only a sub-region of the input data are fully connected with each node. This greatly reduces the number of weights that need to be learnt by the network. The output of the convolutional layer is called a feature map and it is computed as the dot product of the sub-region input and the weights vector. The learnable weight vector is also known as a filter and it allows a CNN to capture useful local features. The Pooling layer can be seen as a form of non-linear down-sampling. It works by dividing the input data into non-overlapping sub-regions and for each sub region a single value is calculated. Two of the most commonly used pooling layers are max-pooling and average-pooling. As the name suggest, max-pooling works by selecting the maximum value in the sub-region as the final output while average-pooling computes the mean value of the sub-region as the final output. Figure 2.20 shows the structure of a CNN with 2D convolutional layers and pooling layers. In this example, the input is a gray-scale image. C1 is a convolutional layer with 4 filters which results in 4 feature maps. S1 is a pooling layer that is used to down-sample the feature maps. C2 and S2 are a further convolutional layer and pooling layer. S2 is then fully connected with the next hidden layer to select the useful local features learned in previous layers.

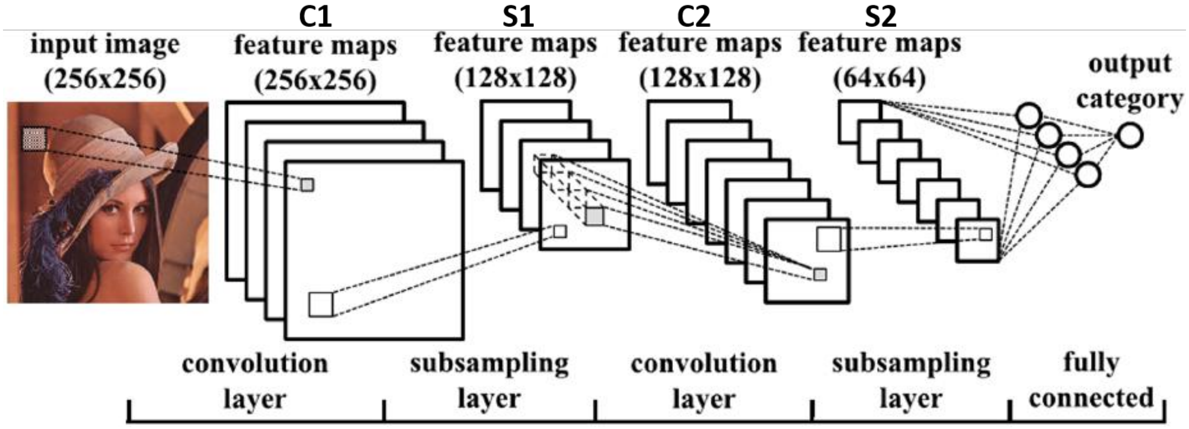


Figure 2.20: Convolutional Neural Network (Cong and Xiao, 2014)

The aforementioned neural network structures such as MLP and CNN are usually called feed-forward networks since the signals can only flow in one direction; *e.g.,* from input to output as indicated by the arrows in Figure 2.19. A Recurrent network is another type of neural network which has directed cycles in the structure as shown in Figure 2.21. This means that by following the direction of the arrow, the signals could flow back to the neuron from which it started. Recurrent neural networks provide a natural way to model sequential data. Each neuron in the hidden layer can be thought of as a deep network in time (see Figure 2.21) thus at each time step t the states of the neuron s can be used to determine the states of the neuron in next time step *t+1.* More specifically an RNN takes the internal state information as an additional input, which ideally consists of all relevant information from the past states of the network. This extends the network's ability to capture temporal information and enhances the learning capabilities to predict the output in time sequence data. Another advantage of RNNs over feed-forward networks is the ability to process arbitrary lengths of input data by using its internal memory and this makes it very popular in speech recognition and language modeling. In order to train an RNN, the *backpropagation through time* (BPTT) algorithm can be used. The key difference of this algorithm compared to the one used in feed-forward neural networks is that the gradient at each layer is computed as the sum of gradients at each time step. One limitation of the original RNNs trained with BPTT is that it is unable to model long-term dependencies since the error flowing backward in time either increases exponentially or vanishes. To solve this problem another network structure called Long Short Term Memory (LSTM) was proposed.

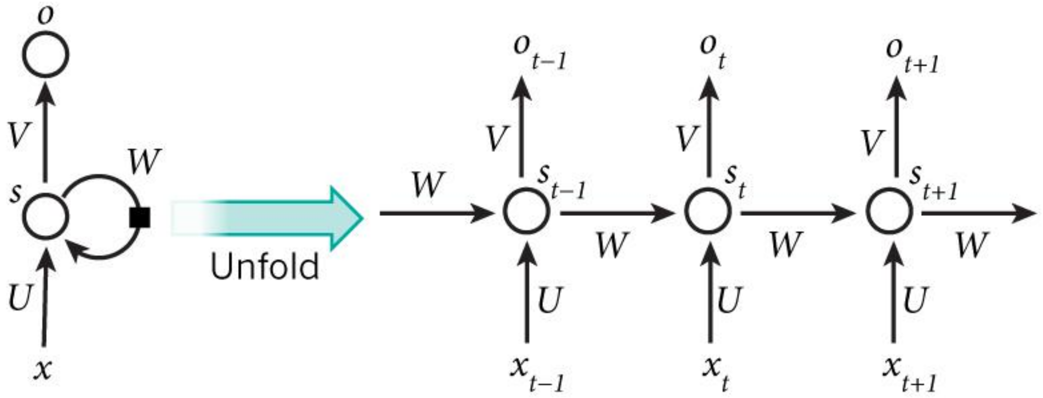


Figure 2.21: An RNN and its unfolded representation. x: input, o:output, V, W, U: weight matrix

LSTM is a type of RNN which was first introduced by Hochreiter and Schmidhuber (1997). The LSTM deals with the vanishing gradient problem by introducing multiplicative gate units which learn to open and close access to the error flow. Gates are a way to control the information flow. A simple LSTM layer includes three types of gate: input gate, forget gate and output gate. Each gate consists of a *sigma* function which has an output between 0 and 1. When multiplying them by another vector, one could decide on how much of that vector to keep. A value of 0 means “do not keep anything”, while a value of 1 means “keep everything”.

2.4.3 Multi-Modal Fusion

As discussed in Section 2.3.2, a human's affective state can be interpreted from different modalities. When multiple modalities are used it is necessary to fuse the results from different modalities to generate one final prediction. There are two commonly used methods to fuse the data: feature-level fusion (early fusion) and decision-level fusion (late fusion). As the name suggests, feature-level fusion takes features from different modalities and concatenates them together to train the model, whereas decision-level fusion first uses features from different modalities to train the model separately and then the outputs from each model are fused together to train a final model.

2.4.4 Performance Evaluation

When evaluating the performance of an affect recognition system, different metrics could be used depending on how the data is annotated or the on nature of the learning task (classification vs. regression).

For data annotated using emotion categories and discrete dimensional labels, the measures of detection rate and F1 score are commonly used (Gunes and Schuller, 2013). The detection rate can be computed at the instance-level (frame-level for visual-based detection and unit-level for vocal-visual-based detection) or segment-level (fixed or variable time interval). The detection rate is calculated as the fraction of the number of correctly detected instances or segments per emotion category divided by the total number of segments for that emotion category. The F1 score also known as F-score or F-measure gives a measure of a classifier's accuracy. It uses both *precision* and *recall* for calculation. Precision is computed as the number of true positives (when both prediction and ground truth are positive) divided by the number of true positives plus the number of false positives (when prediction is positive but ground truth is negative). Recall is defined as the number of true positives divided by the number of true positives plus the number of false negatives (when prediction is negative but ground truth is positive). The F1 score is then calculated as the harmonic mean of precision and recall (Equation. 2.1) with the highest value of 1 and lowest value of 0.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.1}$$

When it comes to evaluating continuously annotated affective state, the optimal evaluation metrics remains an open research issue (Gunes, 2010). The most commonly used metrics are Root Mean Squared Error (RMSE) and Correlation Coefficient (CC) (Grimm and Kroschel, 2005; Wöllmer *et al.*, 2008; Schuller *et al.*, 2012; Valstar *et al.*, 2013, 2014). RMSE is defined as:

$$\begin{aligned} RMSE &= \sqrt{MSE} \\ MSE &= \frac{1}{n}\sum_i (\hat{y}_i - y_i)^2 \end{aligned} \tag{2.2}$$

where n is the total number of samples that need to be predicted, $\hat{y_i}$ is the predicted value of the ith sample, and y_i is the ground truth of the ith sample. RMSE measures the absolute difference between the prediction and ground truth. When RMSE is 0 it means the predicted value matches the ground truth exactly. Correlation Coefficient (CC) is also known as Pearson's product-moment correlation coefficient. It is commonly represented by the Greek letter ρ. CC measures the strength and direction of the linear relationship between two variables. CC is defined as:

$$\rho_{\hat{Y},Y} = \frac{COV(\hat{Y}, Y)}{\sigma_{\hat{Y}}\sigma_Y} = \frac{E\left[\left(\hat{Y} - \mu_{\hat{Y}}\right)(Y - \mu_Y)\right]}{\sigma_{\hat{Y}}\sigma_Y} \tag{2.3}$$

where Y and $\hat{Y}$ are two sequences of ground truth and predictions along time (time series), COV denotes the covariance, σ_Y denotes the standard deviation of Y, E is the expectation and μ_Y is the mean of Y. Correlation coefficients range from -1 to 1.

A value of 1 implies that a strong positive linear relationship exists between $\hat{Y}$ and Y, i.e., as $\hat{Y}$ increases, Y also increases. A value of -1 means a strong negative linear relationship exists where $\hat{Y}$ increases as Y decreases. A value of 0 means that there is no linear relationship between the two variables. A CC score only measures the linear relationship between two series and doesn't measure the difference between them. For instance, if for every prediction the value is 2 times bigger than the actual value, the CC will be 1. When evaluating continuous affect recognition systems it is important to make sure the prediction is close to the ground truth while linearly correlated with the ground truth. More recently the Concordance Correlation Coefficient (CCC) (Ringeval *et al.*, 2015b) has become a popular metric for evaluating continues affect prediction as it combines the Pearson's correlation coefficient with the square difference between the mean of the two compared variables. CCC was first proposed by Lawrence and Lin (1989). It is defined as:

$$CCC = \frac{2\rho_{\hat{Y},Y}\sigma_{\hat{Y}}\sigma_{Y}}{\sigma_{\hat{Y}}^2 + \sigma_{Y}^2 - (\mu_{\hat{Y}} - \mu_{Y})^2} \tag{2.4}$$

where $\rho_{\hat{Y},Y}$ is the correlation coefficient of two time series as defined in equation 2.3. $\sigma_{\hat{Y}}^2$ and σ_{Y}^2 are the variance of prediction and ground truth time series respectively, and $\mu_{\hat{Y}}$ and μ_{Y} are the mean value of each time series. As it can be seen from the equation, unlike CC, predictions that are well correlated with ground truth (high ρ value) but different in value are penalised, resulting in a low CCC score.

2.4.5 Discussion

This section presents the technical background necessary to understand the research performed in this thesis to investigate automatic affect recognition. A number of key decisions are taken based on this state-of-the-art review. Firstly, among different 2D visual features, histogram-based appearance features have been widely employed for continuous affect recognition systems and have achieved good recognition results.

However, there is still little research that investigates the performance of different histogram-based features across different datasets and thus it is unclear if one histogram-based set of features will consistently perform better than another one when using different datasets. The research carried out in this thesis thoroughly investigates the performance of different histogram-based features using different datasets.

Secondly, features used in most of the existing affect recognition systems are extracted from the whole face, and to our best of our knowledge no research has been carried out to study the use of facial parts (*e.g.,* eyes and mouth) for continuous affect recognition. The work described in this thesis involved design of experiments to study if individual facial parts can be used for this purpose.

Thirdly, most of the existing research has focused on using 3D visual features to predict discrete categorical emotions, and no research have been done on using them for continuous affect recognition. In the work described in this thesis, the histogram-based features were applied to the depth map to investigate the performance of 3D features in continuous affect recognition.

For the experiments decribed in this thesis, a variety of machine learning techniques including SVR, CNN and LSTM were selected for evaluation. Decision-level fusion was chosen as the fusion method for two main reasons. Firstly, feature-level fusion could result in high dimensional feature vectors, which usually results in an over-fitted model that does not generalised well for unseen data. Secondly, decision-level fusion could provide greater flexibility in modelling. since different machine learning techniques can be applied to the fusion stage. Finally, in order to compare the results with the state-of-the-art affect recognition systems, three metrics including RMSE, CC and CCC were used.

2.5 Baseline Systems

In order to compare the experimental results reported in this thesis to state-of-the-art, it is important to select a number of continuous affect recognition systems as baseline systems. Since experiments described in this thesis use only the visual modality, only results using visual features are chosen for comparison. In total, four representative baseline systems have been identified. They are the baseline system used for AVEC 2012 (Schuller *et al.*, 2012) and AVEC 2015 (Ringeval *et al.*, 2015b) and the winning system of each challenge (Nicolle *et al.*, 2012; He *et al.*, 2015). At the beginning of this research, it was proposed to implement the AVEC 2012 winning system (Nicolaou *et al.*, 2011) as the only baseline system and apply it to different datasets to compare with the experimental results presented in this thesis. At the time this system was chosen for two reasons. Firstly, it achieved the best average prediction results compared to other entries. Secondly, part of the feature extraction code used in this system is publicly available. Thirdly, by implementing the baseline system one can gain more insight on how the system works. As a result, this system was considered the best performing and most reproducible baseline system. Full implementation details are provided in Appendices A. However, during the implementation of the baseline system it was found that it was not possible to reproduce the results reported in the original paper since there are parameters that are not specified in the paper and details of certain steps are missing. For this reason, the best performed system in AVEC 2015 (He *et al.*, 2015) was also chosen in order to compare the experiments across different datasets. The two baseline systems of AVEC 2012 and 2015 are also included for comparison purposes. As introduced earlier, the AVEC 2012 uses a subset of the SEMAINE dataset which include continuous annotations on arousal, valence, expectation and power dimensions while AVEC 2015 uses a subset of the RECOLLA dataset which only includes continuous annotations on arousal and valence dimensions, thus only arousal and

Table 2.3: State-of-the-art recognition results in terms of CC score on the development partition of different datasets.

System	Dataset	Visual Feature	Learning Method	Arousal	Valence
Schuller *et al.* (2012)	SEMAINE	LBP	SVR	0.151	0.207
Nicolle *et al.* (2012)	SEMAINE	Geometric	Kernel Regression	0.538	0.319
Nicolle *et al.* (2012)	SEMAINE	Global Appearance	Kernel Regression	0.498	0.281
Nicolle *et al.* (2012)	SEMAINE	Local Appearance	Kernel Regression	0.470	0.354
Ringeval *et al.* (2015b)	RECOLA	LGBP-TOP	SVR + LSTM	0.183	0.358
He *et al.* (2015)	RECOLA	LGBP-TOP	LSTM	0.399	0.501
He *et al.* (2015)	RECOLA	LPQ-TOP	LSTM	0.665	0.399

valence dimensions are chosen for the experiments carried out in this thesis. Table 2.3 shows the comparison between the different baseline systems in terms of datasets used, visual features, machine learning techniques and recognition results. There are two things that should be noted. Firstly, the results shown are evaluated using the development partition of both challenges since the labels for the testing partition are not publicly available for evaluation. Secondly, the recognition results are reported in terms of CC score since CCC score was not used for AVEC 2012.

2.6 Conclusion

In this chapter, existing research and the theoretical background related to this work is outlined and used at varying points throughout the following chapters in this thesis. Different emotion models are discussed in Section 2.2. The dimensional model was chosen for the work describe in this thesis due to its ability to encode small differences in affect over time while remaining easy to implement.

In Section 2.3 some of the most commonly used affect datasets in the literature in terms of elicitation method, modalities, emotion models and annotations are reviewed. A problem identified with the existing datasets is that none of them consist of recordings of spontaneous behaviours with 3D information that are also annotated continuously. To address this important gap, it was decided to capture a multi-modal multi-speaker 3D spontaneous affect dataset. More details on how this dataset was constructed are presented in Chapter 3. In addition to the newly captured dataset, the SEMAINE dataset used in AVEC 2012 and the RECOLA dataset used in AVEC 2015 were also selected to evaluate the experimental results obtained from the system described in this thesis.

Various enabling technical components of affect recognition are introduced in Section 2.4 including feature extraction, machine learning, multi-modal fusion and performance evaluation. After examining the current state-of-the-art research, it was decided that the work described in this thesis would focus on the following areas: i) a thorough investigation in the performance of different histogram-based features using different datasets; ii) a study whether individual facial parts can be used for continuous affect recognition; iii) a study of the performance of 3D features in continuous affect recognition by extending the histogram-based features to the depth map; iv) development of a novel system that combines colour and depth and an investigation of the performance gain achieved.

Finally, a number of baseline systems were chosen in order to compare the experimental results obtained by the system described in this thesis to the state-of-the-art.

Chapter 3

Multi-Modal Dataset Collection and Annotation

3.1 Introduction

It is believed that the study of the complex, affective state displayed by humans during social interactions requires rich sets of labelled data that occur naturally in daily-life (Grimm *et al.*, 2008). Such datasets enable researchers to study the relationship between different behavioural cues, e.g., facial expressions and head gestures and their communicative functions during social interactions, e.g., agreement and disagreement. Although there is growing interest in collecting data of social interactions as discussed in Section 2.3.4, to our best of knowledge, there is still no dataset that includes recordings of spontaneous behaviours with both audio-visual and depth data that is also annotated continuously in a multi-dimensional affective space. This Chapter introduces and describes the creation of one of the first spontaneous and continuously annotated multi-modal dataset focused on human interaction during a debate. The novelties of our multi-modal dataset are as follows:

(i) it is based on a three-way debate scenario, allowing researchers to study the spontaneous affective state in relation to different modalities and to study affective response between different participants.

(ii) it contains multimodal data including video, depth and audio modalities with detailed continuous annotation on different dimensions.

The remainder of this chapter is organised as follows: Section 3.2 explains how the dataset was constructed. Section 3.3 introduces a multi-modal capture platform designed for this data capture. Section 3.4 details the segmentation and annotation process of the dataset. Finally Section 3.5 describes a statistical analysis on the dataset to validate its usefulness as a research tool.

3.2 Dataset Construction

3.2.1 Participants and Environment

In total 16 participants from Dublin City University and Bell Labs Ireland were recruited for the dataset capture. The 14 participates were 4 females and 12 males with age group ranging from 20 to 50 years old. Six offices with various background and illuminations were used during the capture. In order to capture both facial expressions and upper-body gestures, each participant was arranged to sit one meter away from the screen.

3.2.2 Procedure

At the beginning of each capture session, the 3 participants were introduced to each other, followed by an introduction to the experiment. Then they were separated in three different offices and a debate topic was given. Similar to Mahmoud *et al.* (2011), a wizard-of-oz method was used. Participants were told at the beginning of the experiment that their video and audio would be recorded for face and voice recognition purposes. Not knowing the real objectives of the experiment avoided having participants exaggerate or mask their true affect state (Mahmoud *et al.*, 2011). Each capture session ended when either a time limit was reached (60 minutes) or the debate came to a natural conclusion.

3.2.3 Elicitation of Affective States

As discussed in Section 2.3.1, there are three main types of interaction behaviour for eliciting affective state: (i) posed behaviour. (ii) induced behaviour and (iii) spontaneous behaviour. Since this dataset focuses on collecting spontaneous behaviour in a real-life situation, a debate scenario was chosen. Compared to other scenarios, a debate scenario features the following attributes: (i) debate occurs naturally in everyday life, such as in a meeting, when watching a football match, or watching movies, and participants are typically moved by real motivations leading to highly spontaneous affective states; ii) debate scenarios convey rich affective state and social behaviours such as conflicts, dominance, agreement/disagreement and interest/non-interest (Vinciarelli *et al.*, 2009).

To start the debate, the following topics were selected:

1. How Ireland performed in the Six Nations Rugby match (a high profile sporting event).
2. Should Ireland reduce the minimal wage?
3. Will the Irish economy take off in the future?
4. Do humans have free will?
5. Do humans have a moral obligation to be vegetarian?

The first topic was used in two capture sessions. The first capture session consists of three sports fans, allowing the capture of strong interest. The second capture session includes two sports fans and one non-sports fan. This allows the capture of

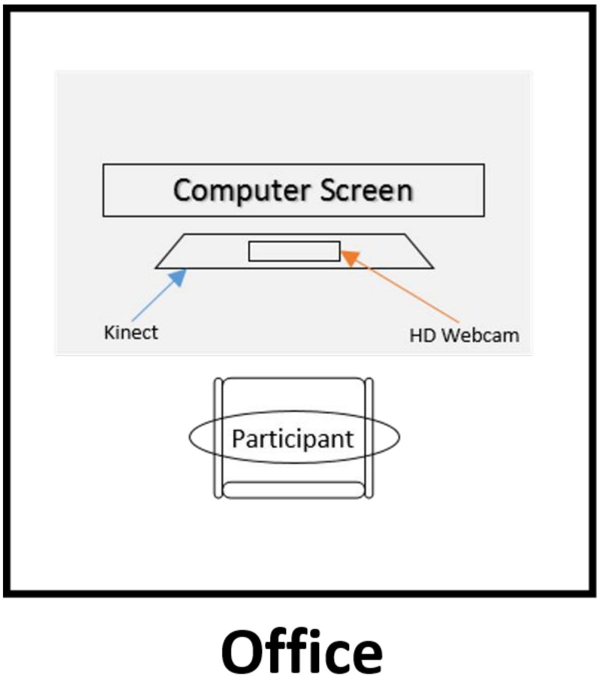


Figure 3.1: Plan-view of capture environment layout

rich interest and non-interest. The rest of the topics were used to enable the capture of agreement/disagreement and positive/negative valence.

3.3 Multi-Modal Recording Setup

3.3.1 Sensors and Software

The capture environment setup is shown in Figure 3.1 and was replicated in each office. A High-Defintion (HD) webcam (Logitech C910) was used and fixed on top of the Microsoft Kinect to collect the visual signals. The microphones in the HD webcam were used to capture the audio signals and the Kinect was used to capture the depth information. Two computers were used in each office, one computer was used by the participant to communicate using Google Hangout, while the other computer was used to capture the multi-modal data. The HD webcam provided 1280×720 resolution colour images at 30 frames per second. The Kinect sensor consists of a normal RGB camera and an infrared camera. The RGB camera is able to provide 640×480 color image and the infrared camera is used to capture structured light and calculate a 640×480 11-bit disparity map. A headphone was used by each participant to prevent capturing other participants' voice. In order to reduce the load on the hard drive, only the depth stream from the Kinect was recorded. The audio was recorded using the microphone on the HD webcam at 16 bit and 96kHz. Camera calibration (Zhang, 1999) was performed between the HD webcam and Kinect infrared camera in order to map the depth information to the RGB image.

To capture the video, audio and depth simultaneously for each participant, modifications were made to the open source video capture software Virtualdub (Lee, 2013) to support reading depth streams from the Microsoft Kinect. The video stream was compressed using MJPEG, while the depth stream was saved in the ONI format developed and used by the OpenNI framework. Given that participants are sited in different locations, a program was developed to start and stop recording at the same time across different locations through asynchronous TCP socket communication. On each local machine an AutoIT Script was created to automate the capture process. An overview of the multi-station capture system is shown in Figure 3.2. Figure 3.3 shows sample data from the captured dataset with different arousal and valence values. It can also be seen that the captured dataset consists of various background and lighting conditions.

3.4 Annotations

3.4.1 Segmentation

Since each debate section usually lasted from 40 to 60 minutes, the videos were segmented into 5 to 10 minute clips for easier annotation. After watching different

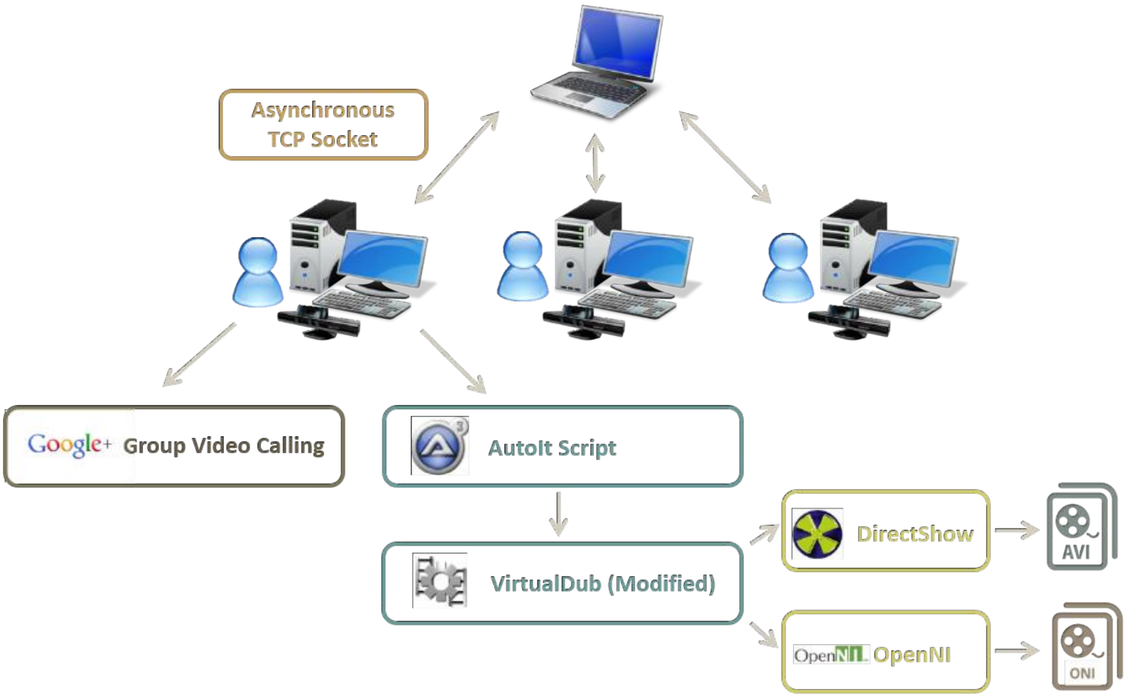


Figure 3.2: Overview of the Multi-Station Capture System

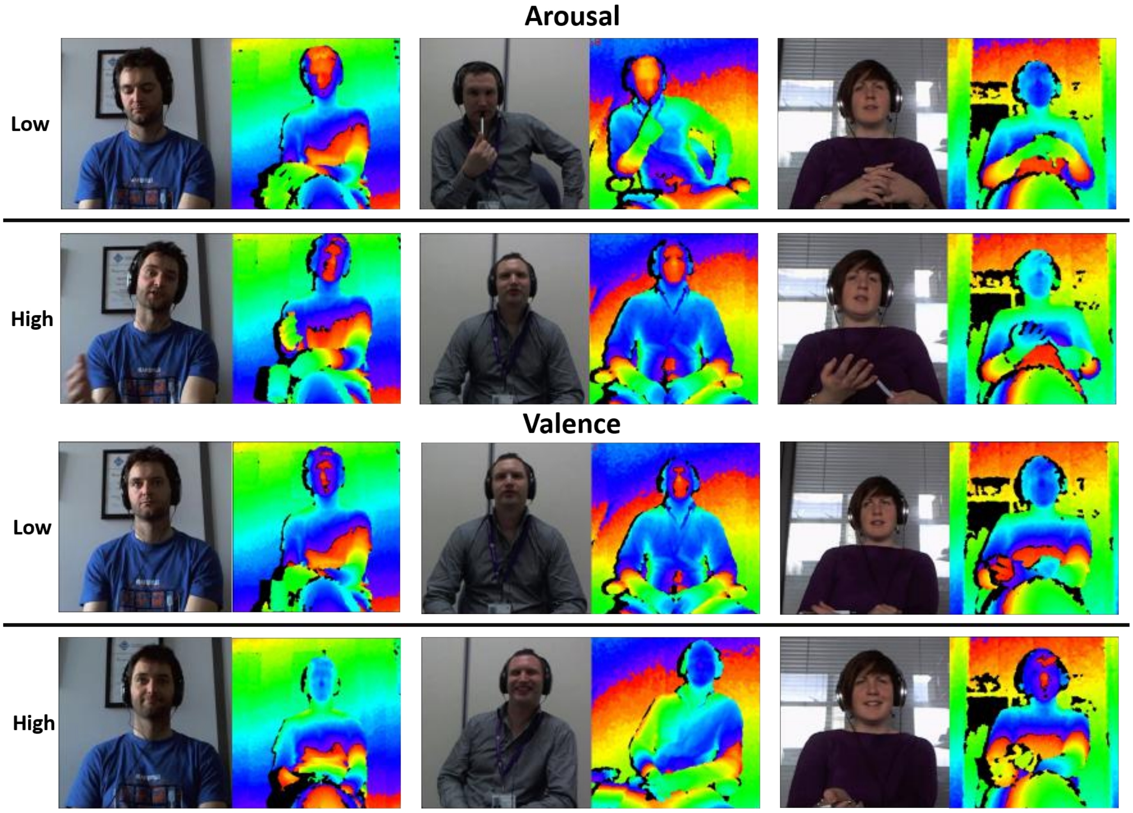


Figure 3.3: Sample data from the dataset

capture sessions, it was found that the beginning of a session usually consists of warm up chat while at the end of a session the participant might end up discussing other topics. These two parts do not involve as many different affective states compared to the middle part. As a result only the middle part of each session was annotated. This resulted in 36 video clips consisting of approximate 5 hours and 30 minutes of data.

3.4.2 Annotation Guidelines

Three independent annotators were hired. Before the annotation task, each annotator was introduced to the annotation task. Then they were required to complete a set of training tasks to test their affect recognition skill and to become familiar with the use of Gtrace (See Section 2.3.3). The first training task involved the identification of emotions expressed on the face. The second task required participants to describe the emotional state shown in a video clip. The third task involves mapping a list of 24 emotional keywords to a valence-arousal 2-dimensional space. Task 4 involved annotating a list of sample videos from the SEMAINE dataset (McKeown *et al.*, 2012) using Gtrace. Upon completion of the training tasks and having become familiar with the Gtrace annotation tool, the annotators were prompted to start the annotation tasks. For convenience, annotators were allowed to pause or restart an annotation at a any given time. To help the annotator better follow the conversation, audio clips from each participant during the same capture session were mixed together. For each video clip, five dimensions were continuously annotated, including arousal, valence, agreement, interest and content. The value range for all dimensions was set to $[-1, 1]$.

3.4.3 Post-processing

The annotations were first post-processed to remove duplicated annotations, and cropped to be temporally aligned with the video sequences. The annotation data was then binned with a frame rate fixed to match the video frame rate, which is a 33ms duration bin in our case.

Because the video and depth signal are captured from different sensors and there was no hardware synchronization available, subsequent manual synchronisation was required. Although the video and depth frame programmatically start at the same time, the Kinect usually takes a longer time for the first frame to arrive and the delay is not constant and may depend on the computer hardware specification. To deal with this problem, after each capture session was segmented into clips as described in the previous section, the first frame for each depth video clip was manually aligned to its corresponding colour video clip. This eliminated the delay between the first frame of the colour and depth streams.

3.5 Statistical Analysis

To evaluate the reliability of the annotations, three metrics are used, including the percentage of positive frames, the mean Correlation Coefficient (CC) and Cronbach's α. The percentage of positive frames is defined as the number of frames that have annotations greater than zero, divided by the total number of frames. It is an indication of whether the collected data is balanced, where ideally 50% positive frames and 50% negative frames as desired. The mean correlation coefficient measures the average linear relationship between different annotations. It is calculated using Equation 2.3. Cronbach's α is used to measure the internal consistency. It is defined as:

$$\alpha = \frac{K\bar{c}}{(\bar{v} + (K-1)\bar{c})} \tag{3.1}$$

where K is total number of annotators, $\bar{v}$ is the average variance of each annotation and $\bar{c}$ is the average of the covariances matrix between different annotations excluding the current annotation. The standard description of α levels is (George and Mallery, 2003):

- > 0.9 Excellent
- > 0.8 Good
- > 0.7 Acceptable
- > 0.6 Questionable
- > 0.5 Poor
- < 0.5 Unacceptable

It should be noted that 0.6 is the lowest value commonly considered acceptable in practice.

The results of the statistical analysis of the proposed dataset are shown in Table 3.1. The analysis of the raw data shows an extremely high percentage of positive arousal and interest, and a high percentage of positive valence, agreement and content. This could be explained by the nature of the debate scenario since during a debate, participants are usually highly engaged in the conversation. Since the annotation itself is a subjective measurement (e.g., different annotators could give different annotations for the same affective state), Zero mean normalisation as employed in Ringeval *et al.* (2013) can be applied to balance the data. In terms of mean correlation coefficients, the analysis shows good correlations between annotators on the valence, agreement and content dimensions, and high correlations on the arousal and interest dimensions. The analysis also shows good internal consistency for the arousal and interest dimensions, and acceptable consistency for valence, agreement and content dimensions. One explanation of the low internal consistency is that

Table 3.1: Statistics of the annotations

Statics Properties	**Arousal**	**Valence**	**Agreement**	**Content**	**Interest**
% Pos Frame	97.3	73.3	79.6	74.8	94.6
Mean Corr.	0.76	0.47	0.46	0.39	0.66
Mean α	0.89	0.66	0.63	0.60	0.83

there are situations where the signs of affective state on these dimensions are ambiguous (e.g., a person looks happy but his/her voice sounds unhappy). To further investigate this, the annotations from different annotators were played back along with the video recordings. It was found that during the debate, there are situations where no clear indications of participant's affective state on valence dimension are presented and the the annotators tended to interpret the affective state differently. Figure 3.4, 3.5, 3.6 and 3.7 shows the annotation on arousal, valence, agreement and content dimensions for the same segment of a video recording. As it can be seen, the annotators gave very different values for valence (between -0.3 to 0.4), agreement (between 0 to 0.7) and content (between 0.2 to 0.6) dimensions. In contrast, the annotators gave very similar values (between 0.8 to 0.9) for the arousal dimension.

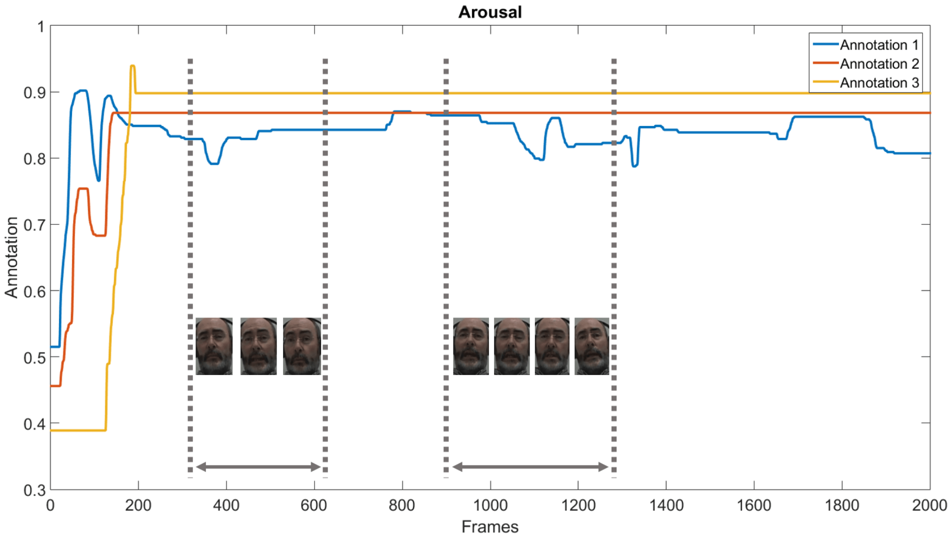


Figure 3.4: Annotation comparison for different annotators on arousal dimension

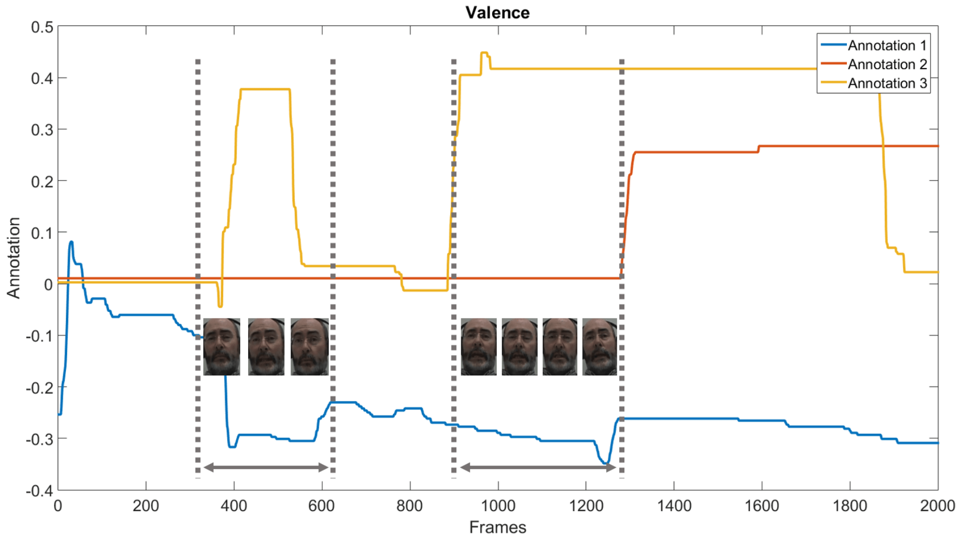


Figure 3.5: Annotation comparison for different annotators on valence dimension

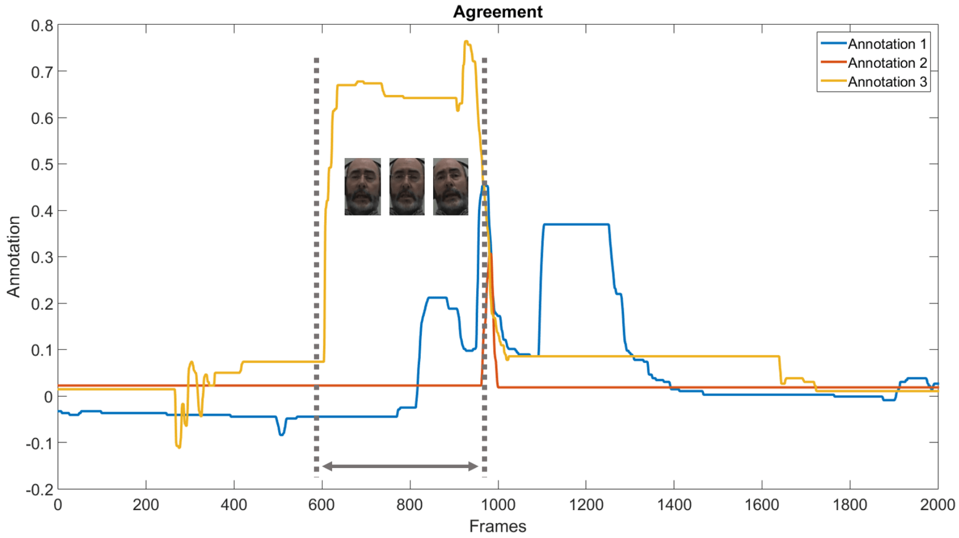


Figure 3.6: Annotation comparison for different annotators on agreement dimension

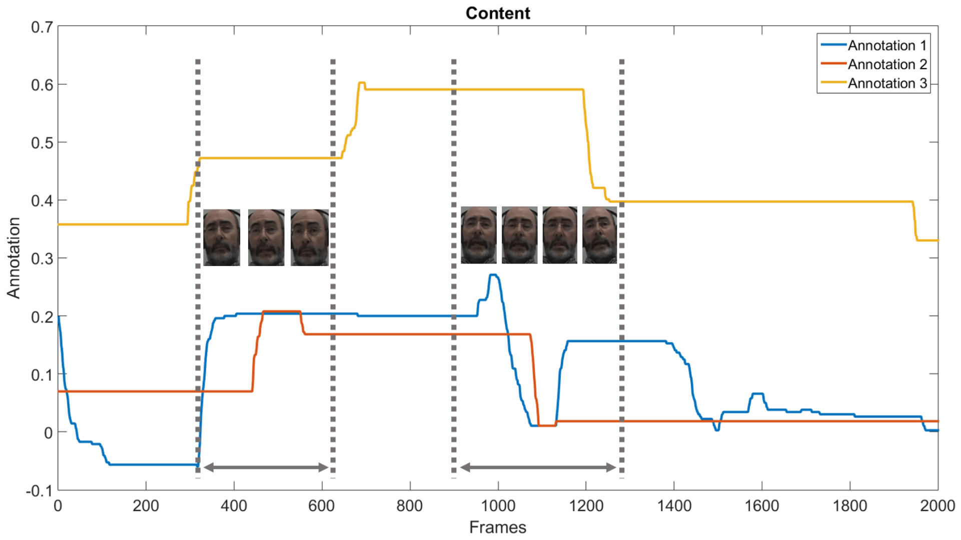


Figure 3.7: Annotation comparison for different annotators on content dimension

3.6 Conclusion

In this chapter, the development of a data capture platform capable of collecting a synchronised colour and depth stream is first described and then it is described how this was used to create a multi-modal dataset in real-world settings. To the best of our knowledge, this is one of the first spontaneous and continuously annotated multi-modal datasets based on human interaction during a debate. 16 participants were recorded during a sequence of debates in a three-way video conference setup. Recordings include video signals, audio signals and depth signals. In total, over five hours of data have been manually annotated in five dimensions including arousal, valence, agreement, content and interest. Due to the nature of the debate scenario, participants are usually highly engaged in the conversation, which meant that the percentage of positive frames for the arousal and interest dimension are significant higher. The Cronbach alpha measure shows correlation on arousal and interest dimensions and acceptable correlation on valence and content dimensions, making it a suitable dataset for the work proposed in this thesis. After plotting the annotation along with the video, it was found that the low Cronbach's measure was mainly caused by situations where no clear indications of participants affective state for valence, agreement and content dimension are presented and the annotators tended to interpret the affective state differently. One way to solve this problem could be to increase the number of annotators since more annotators could give more reliable annotations. However, this would further exacerbate an already extremely tedious manual process.

Chapter 4

Affect Recognition using Colour Video Data

4.1 Introduction

As discussed in Section 2.4.1.1, facial features extracted from colour video data can be divided into two main categories: geometric features and appearance features. Geometric features are robust to lighting conditions since the focus is on coordinates of landmarks rather than the intensity of the pixel. However, they are very sensitive to facial landmarks registration errors as they are calculated purely based on the landmark coordinates. Compared to geometric features, appearance features are also robust to illumination variations, as they are extracted from small regions, and usually do not depend on facial landmark registration. Among different appearance features, the histogram-based appearance features have been used widely in affect recognition since they can be normalised to increase the robustness of the overall representation and are computationally simple when compared to filter-bank-based features. Some of the most commonly used histogram-based features in affect recognition include Local Binary Pattern (LBP), Local Phase Quantisation (LPQ), Local Gabor Binary Patterns (LGBP), Local Binary Patterns on Three Orthogonal Planes

Table 4.1: Histogram-based features and corresponding datasets that used to evaluate their performance

Feature Type	**Dataset**
LBP	SEMAINE
LPQ	AViD
LGBP	-
LBP-TOP	RECOLA
LPQ-TOP	RECOLA
LGBP-TOP	AViD, RECOLA

(LBP-TOP), Local Phase Quantisation on Three Orthogonal Planes (LPQ-TOP) and Local Gabor Binary Patterns on Three Orthogonal Planes (LGBP-TOP). For instance, the LBP feature was used in systems developed by Schuller *et al.* (2011, 2012), Van Der Maaten (2012), Savran *et al.* (2012a) and Sánchez-Lozano *et al.* (2013). The LPQ feature was employed in the work carried out by Valstar *et al.* (2013) and Kaya *et al.* (2014). Kächele *et al.* (2015) compared the performance of the LBP-TOP feature with Histograms-of-oriented-gradients (HOG) feature and Pyramids of histograms of oriented gradients in three orthogonal planes (PHOG-TOP) feature. The work carried out by He *et al.* (2015) investigated the performance of the LPQ-TOP feature. The LGBP-TOP feature was adopted in systems developed by Valstar *et al.* (2014), Senoussaoui *et al.* (2014), Ringeval *et al.* (2015b), He *et al.* (2015) and Chen and Jin (2015). Although the performance of different histogram-based features has been studied, their performance are usually evaluated using different datasets as shown in Table 4.1 and different learning techniques. Thus there is no clear indication of which histogram-based feature performs best for continuous affect recognition. Furthermore, current research uses fixed configurations when extracting histogram-based features, and it is not clear how different configurations will affect the recognition result. No current research has reported the performance of LGBP features for continuous affect recognition.

The Convolutional Neural Network (CNN) introduced in Section 2.4.2.2 has been applied to solve a variety of problems such as image recognition, video analysis and natural language processing. More recently, it has also been applied for continuous affect recognition. For instance, Chao *et al.* (2015) explored the use of CNN as a feature extractor to extract appearance features for continuous affect recognition. The AlexNet CNN structure (Krizhevsky *et al.*, 2012) was pre-trained using a combination of Celberity Faces in the Wild (CFW) (Zhang *et al.*, 2012) and FaceScrub dataset (Ng and Winkler, 2014). The outputs of the last convolutional layer are used as face features. It achieved a CC score of 0.348 on the arousal dimension and 0.561 on the valence dimension on the development partition of the AVEC 2015 dataset. The work carried out by Khorrami *et al.* (2016) uses a simple three-layer CNN as a feature extractor, and the network achieved a CC score of 0.554 on valence dimension on the development partition of AVEC 2015 dataset. However, to the best of our knowledge, the current deep learning approaches all use features extracted from the entire face region and none of them have presented the use of individual facial parts such as eyes and mouth for continuous affect recognition.

In this Chapter, different appearance features extracted from the video modality are examined. In particular, experiments have been designed to thoroughly study the performance of different histogram-based features and to investigate if it is possible to use individual facial parts for continuous affect recognition.

4.2 Experiment 1: Configuration of histogram-based features

The aim of Experiment 1 is to investigate the best configurations for different histogram-based features. Specifically, experiments are carried out to identify the best block size for different histogram-based features. More blocks means more detailed description of the input image, which in turn results in higher dimensional features and usually longer feature extraction time. For LBP, LPQ, LBP-TOP, LPQ-TOP features, the code provided by CMV, University of OULU is used [1]. For the LGBP and LGBP-TOP features, the code developed by Michel Valstar is used [2].

In this experiment, the AVEC 2015 dataset is selected for two reasons. Firstly, it has been widely used for continuous affect research. Secondly, it consists of more subtle expression changes when compared to other datasets, as suggested by Kächele *et al.* (2015).

4.2.1 Data Pre-processing

To extract the histogram-based features, 49 facial landmarks are first detected and tracked for each video frame using the Supervised Descent Method (SDM) proposed by Xiong and De la Torre (2013). The face is then extracted using the minimal and maximal coordinates from the tracked facial landmarks for the x-axis and y-axis respectively. Finally, the face image is re-sized to 96×96 for feature extraction. An overview of the pre-processing steps is shown in Figure 4.1.

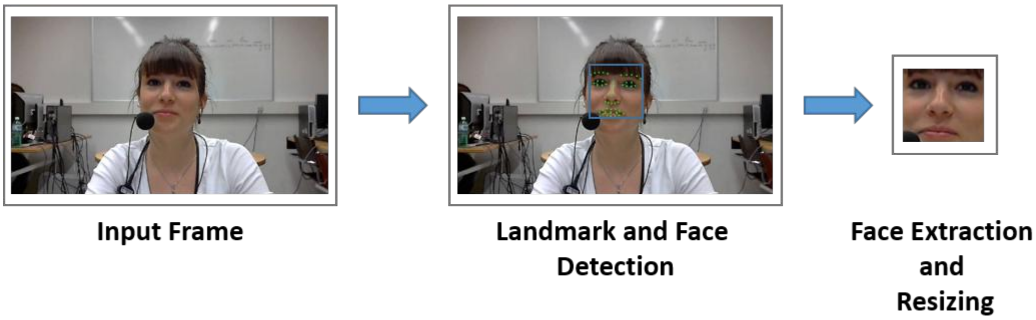


Figure 4.1: Overview of pre-processing steps

[1] http://www.cse.oulu.fi/wsgi/CMV/Downloads

[2] http://www.cs.nott.ac.uk/~pszmv/resources/visum_package.zip

4.2.2 Experimental Procedure

In this experiment, the most commonly used configurations for different histogram-based features in the literature (Schuller *et al.*, 2012; Valstar *et al.*, 2013; Zhao and Pietikainen, 2007; Valstar *et al.*, 2014) are used for comparison purposes. The configurations are shown in Table 4.2. The window size used for Fourier phase computation is set to 7 for LPQ features, and 7 for LPQ-TOP features on *x*-*y* planes and 3 for *x*-*t* and *t*-*t* planes as suggested by Jiang *et al.* (2014). To reduce the feature dimensions of LGBP and LGBP-TOP features, the number of scales is set to 2 while the number of orientations is set to 3. This generates six Gabor filters in total. The block sizes are evaluated at 1×1, 2×2 and 4×4. To further reduce the dimensions, no block size is applied to the temporal axis (e.g. for block size 4×4, only the *x*-*y* plane is divided into 4×4 blocks, *x*-*t* and *y*-*t* planes are divided to 4*1 blocks). Since the feature dimensions for some histogram-based features could be high (shown in Table **??**, which may lead to overfitting, the dimensionality reduction is needed. Similar to the work carried out by Ringeval *et al.* (2015b), the Singular Value Decomposition (SVD) from a low-rank approximation (Halko *et al.*, 2011) is applied to all the extracted features. To ensure the reduced features have similar dimensions, the rank of each block size configuration is set to 1/2, 1/12 and 1/50 of the feature dimensions respectively, and the reduced number of features are selected to cover 98% of the variation of the original features. The original number of features and number of reduced features for each histogram-based feature are shown in Table

Table 4.2: Configuration for different histogram-based features

Feature Type	Neighbor	Radius	Mapping	Time Interval(frames)
LBP	8	1	Uniform	-
LPQ	-	-	-	-
LGBP	8	1	Uniform	-
LBP-TOP	8	1	Uniform	5
LPQ-TOP	-	-	-	5
LGBP-TOP	8	1	Uniform	5

Table 4.3: Original number of features and reduced number of features for different block sizes

	1×1		2×2		4×4	
	Total	Reduced	Total	Reduced	Total	Reduced
LBP	59	25	236	19	944	19
LPQ	256	121	1024	81	4096	76
LGBP	177	79	1416	108	5664	112
LBP-TOP	177	86	708	58	2832	58
LPQ-TOP	708	335	3072	242	12288	237
LGBP-TOP	1416	649	2832	214	8496	152

4.3. Furthermore, for static features (LBP, LPQ and LGBP), the feature values are forced to be zero where a face is missing, while for dynamic features (LBP-TOP, LPQ-TOP and LGBP-TOP), the feature values are forced to be zero if the volume data contains more than one frame where a face is missing. These feature values are excluded from the modelling process since they can be misleading and skew the training process.

For machine learning, a linear Support Vector Machine for regression (SVR) was used to perform the regression task using the liblinear library (Fan *et al.*, 2008). The L2-regularised L2-loss dual solver was chosen and a unit bias was added to the feature vector. During training, feature vectors containing all zeros were dropped. The complexity parameter c of the SVR was optimised in the range of $[10^{-5} - 10^{0}]$ while other parameters were kept as default.

4.2.3 Experimental Results and Analysis

The result for different block size configurations are shown in Table 4.4. The score is calculated in terms of Root Mean Square Error (RMSE: lower is better), Correlation Coefficient (CC: higher is better) and Concordance Correlation Coefficient (CCC: higher is better) as discussed in Section 2.4.4.

To better compare the performance of different histogram-based features and block sizes, bar charts are plotted in terms of CC score, since CCC score can usually

Table 4.4: Results on development partition for different block sizes. The best results in different metrics are highlighted in bold.

	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP	0.190	0.158	0.050	0.127	0.141	0.091
LPQ	0.201	0.191	0.127	0.132	0.280	0.234
LGBP	0.222	0.036	0.017	0.135	0.141	0.094
LBP-TOP	**0.185**	0.295	0.213	0.128	0.185	0.127
LPQ-TOP	0.187	**0.315**	**0.245**	**0.125**	**0.336**	**0.293**
LGBP-TOP	0.196	0.078	0.022	0.137	0.092	0.071

(a) Block size: 1 × 1

	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP	0.191	0.146	0.046	**0.126**	0.242	0.192
LPQ	0.211	0.120	0.095	0.133	0.234	0.206
LGBP	0.228	0.015	0.009	0.149	0.138	0.094
LBP-TOP	**0.188**	**0.259**	0.081	0.134	0.147	0.121
LPQ-TOP	0.197	0.154	**0.105**	0.127	**0.281**	**0.249**
LGBP-TOP	0.202	0.151	0.095	0.136	0.172	0.135

(b) Block size: 2 × 2

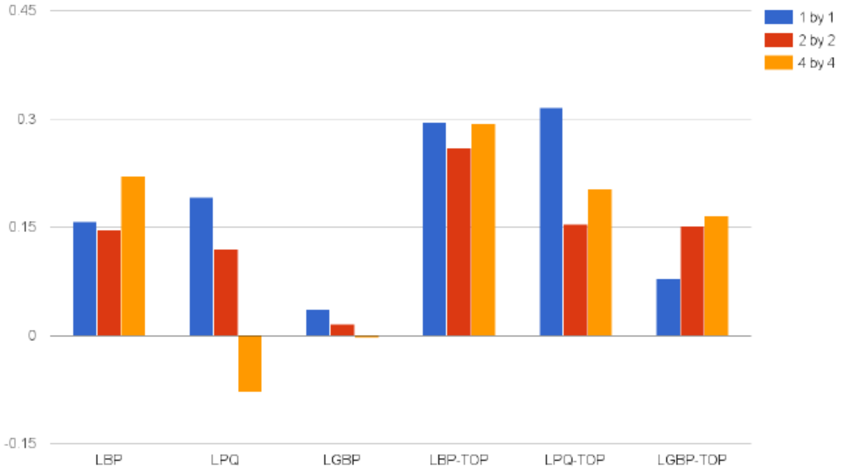
	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP	0.217	0.220	**0.146**	**0.118**	0.302	**0.213**
LPQ	0.206	-0.080	-0.037	0.135	0.228	0.170
LGBP	0.197	-0.003	-0.001	0.123	0.207	0.138
LBP-TOP	**0.189**	**0.293**	0.108	**0.118**	0.310	0.210
LPQ-TOP	0.193	0.202	0.098	0.122	**0.338**	0.166
LGBP-TOP	0.193	0.166	0.045	0.124	0.187	0.088

(c) Block size: 4 × 4

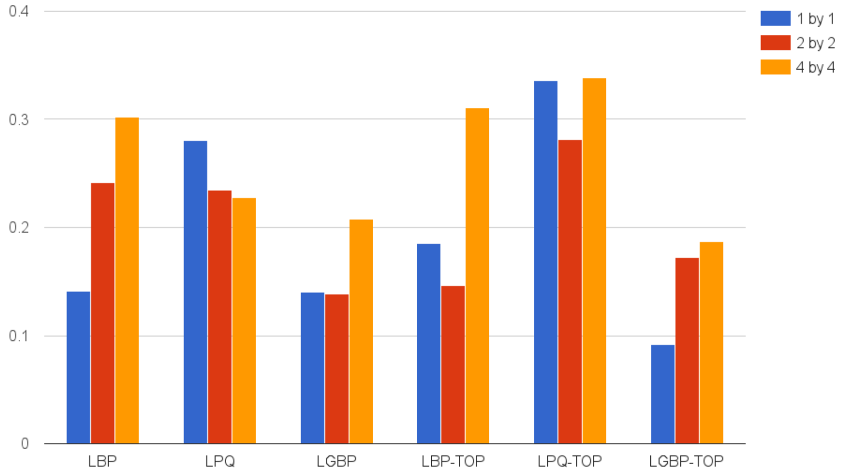
be improved by post-processing as can be seen in later experiments. Figure 4.2 illustrates the recognition results for different histogram-based features in various block size configurations.

Overall, it can be seen that the dynamic features generally achieved better CC scores on both arousal and valence dimensions compared to their static counterparts and this agrees with the findings in Almaev and Valstar (2013) and Jiang *et al.* (2014) for facial action unit detection. For LBP features, increased block size yields better recognition results. The 4 × 4 configuration achieved the best result on both arousal and valence dimensions. For LPQ features, increased block size results in lower CC score for both arousal and valence dimensions. The 1 × 1 block size achieved the best recognition result. For LGBP features, increased block size decreases performance on the arousal dimension, but increases performance on the valence dimension. For LBP-TOP features, the 4 × 4 configuration achieved similar results as the 1 × 1 configuration for arousal dimension, but achieved much better results on the valence dimension when compared to other configurations. In contrast, for LPQ-TOP features, the 1 × 1 configuration achieved the best performance on the arousal dimension and similar results as the 4 × 4 configuration on the valence dimension. Finally, for LGBP-TOP features, the performance increases linearly with the block size, with the 4 × 4 configuration achieved the best result. Among all the features, LPQ-TOP features achieved the best recognition result for both arousal and valence dimensions followed by LBP-TOP features. It is also interesting to note that most features extracted using a 1 × 1 configuration generally perform well on the arousal dimension, while features using a 4 × 4 configuration generally perform well on the valence dimension.

Table 4.5 shows the computation time of the different histogram-based features with different block size configurations. All computation was tested on a Intel Core i7-2600K processor with 16 gigabytes of RAM. As can be seen, increased block size generally increases the computation time except for the LBP-TOP feature. For static



(a) Arousal



(b) Valence

Figure 4.2: Comparison of the performance of different block size configurations in terms of CC score on Arousal and Valence dimensions

Table 4.5: Computation time in seconds for histogram-based features with different block size configurations. The LBP, LPQ and LGBP features are computed per frame. The LBP-TOP, LPQ-TOP and LGBP-TOP features are computed every 5 frames.

	LBP	**LPQ**	**LGBP**	**LBP-TOP**	**LPQ-TOP**	**LGBP-TOP**
1×1	0.048	0.057	0.096	0.187	0.129	0.178
2×2	0.053	0.065	0.218	0.182	0.180	0.280
4×4	0.061	0.074	1.383	0.167	0.279	2.090

features, the LGBP feature require the most computation time while for dynamic features, LGBP-TOP is the most computationally expensive.

4.3 Experiment 2: Annotation Delay

As discussed in Section 2.2.2, one of the main challenges of using a continuously annotated dimensional dataset is the delay in the annotation. To deal with this problem, various approaches have been proposed. For instance, Nicolaou *et al.* (2011) estimated constant shifts between the prediction and ground truth to minimise their mean square error (MSE). Nicolle *et al.* (2012) assumes a linear relationship between the features and labels, and proposed a correlation-based measure to find the delay. Nicolaou *et al.* (2012) introduced the *dynamic probabilistic canonical correlation with time warping (DPCTW)* approach to compensate for local delays between different annotators. The aim of this experiment is to study the performance of different block size configurations after the annotation delay has been compensated and to investigate the general annotation delays for arousal and valence dimension.

4.3.1 Experimental Procedure

Similar to experiment, the AVEC 2015 dataset was used for this experiment. The same experimental setup as in previous experiment was used and the time delay was treated as a hyperparameter. It was optimised by CC score for arousal and valence respectively on the training partition. The delay was considered as a constant shift from 0 to 8 seconds with a step size of 0.4 seconds. Each SVM model was trained with the shifted ground truth and the best results are reported in Table 4.6.

4.3.2 Experimental Results and Analysis

A bar chart is provided to compare performance as shown in Figure 4.3. Overall the 4 × 4 LPQ-TOP feature achieved best results on the arousal dimension while the 4 × 4 LPQ-TOP features perform best on the valence dimension. Compared to the first experiment, the performance on all configurations have been improved. On the arousal dimension, the best CC score is improved by 38% (from 0.315 to 0.435) while on the valence dimension, the best CC score is improved by 31% (from 0.338 to 0.445).

In order to analyse the annotation delay on the arousal and the valence dimensions, the recognition results for different features are plotted in terms of different time delays. This is shown in Figure 4.4, 4.6, 4.5 and 4.7. By looking at the maximum values for each delay curve, one can identify an average delay between 1.6 to 2 seconds for the arousal dimension and 1.6 to 2.4 seconds for the valence dimension across different histogram-based features.

4.4 Experiment 3: The effect of post-processing

Since the predictions from the machine learning step usually suffer from issues such as bias, scaling and noise, various post-processing techniques have been employed for recently developed affect recognition systems. For instance, Kächele *et al.* (2015) scaled the predictions using the minimum and maximum value from the training partitions. Exponential smoothing is used to remove the noise in the work carried out by Chen and Jin (2015). He *et al.* (2015) employed Gaussian smoothing with both fixed and variable window length to remove the noise. The aim of this experiment is to study how post-processing steps can be used to improve the recognition result.

Table 4.6: Results on development partition for different block sizes with shifted annotation. The best results in different metrics is highlighted in bold

	Arousal				**Valence**			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP	1.2	0.189	0.189	0.061	2.4	0.127	0.183	0.130
LPQ	0.8	0.202	0.201	0.143	1.2	0.132	0.334	0.284
LGBP	3.6	0.223	0.123	0.054	2	0.124	0.204	0.105
LBP-TOP	1.6	**0.177**	0.406	0.246	2.4	0.125	0.249	0.183
LPQ-TOP	1.6	0.181	**0.430**	**0.348**	1.6	**0.122**	**0.403**	**0.373**
LGBP-TOP	2	0.193	0.205	0.065	7.2	0.130	0.169	0.134

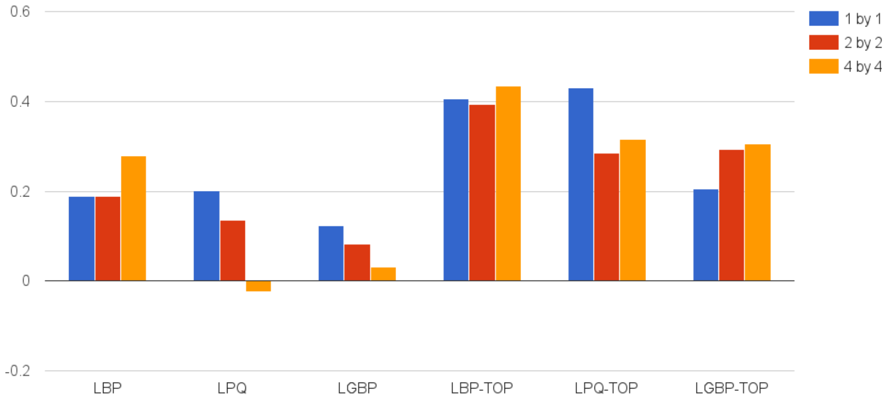
(a) Block size: 1 × 1

	Arousal				**Valence**			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP	1.2	0.189	0.188	0.057	1.6	0.122	0.289	0.231
LPQ	0.8	0.200	0.136	0.090	2	0.124	0.362	0.327
LGBP	4.8	0.224	0.082	0.039	2.4	0.131	0.191	0.065
LBP-TOP	2	**0.179**	**0.393**	**0.227**	1.6	0.132	0.205	0.174
LPQ-TOP	2	0.188	0.284	0.176	1.6	**0.121**	**0.378**	**0.356**
LGBP-TOP	2	0.193	0.293	0.192	1.6	0.135	0.193	0.155

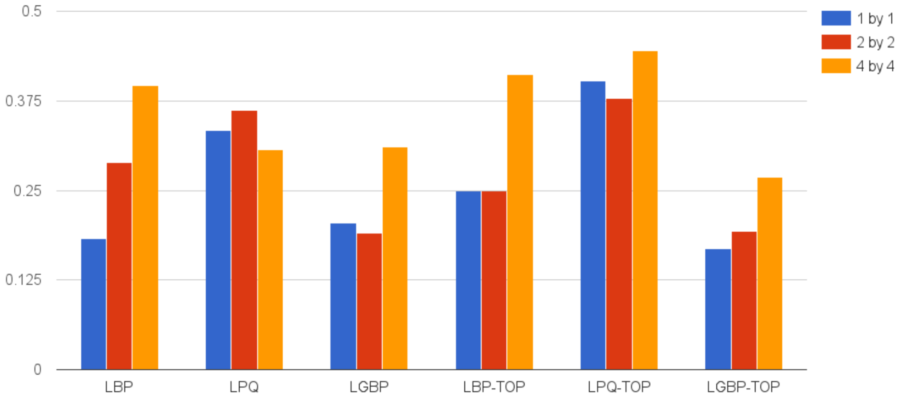
(b) Block size: 2 × 2

	Arousal				**Valence**			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP	1.6	0.200	0.279	0.144	2.4	0.113	0.397	0.304
LPQ	6.4	0.266	-0.024	-0.022	1.2	0.129	0.307	0.240
LGBP	3.2	0.232	0.032	0.018	2	0.129	0.311	0.231
LBP-TOP	2	**0.184**	**0.435**	**0.225**	2.4	0.116	0.412	0.243
LPQ-TOP	1.6	0.188	0.316	0.195	1.6	**0.111**	**0.445**	**0.303**
LGBP-TOP	2	0.189	0.305	0.097	2	0.121	0.269	0.135

(c) Block size: 4 × 4



(a) Arousal



(b) Valence

Figure 4.3: Comparison of the performance of different block size configurations in terms of CC score on Arousal and Valence dimensions

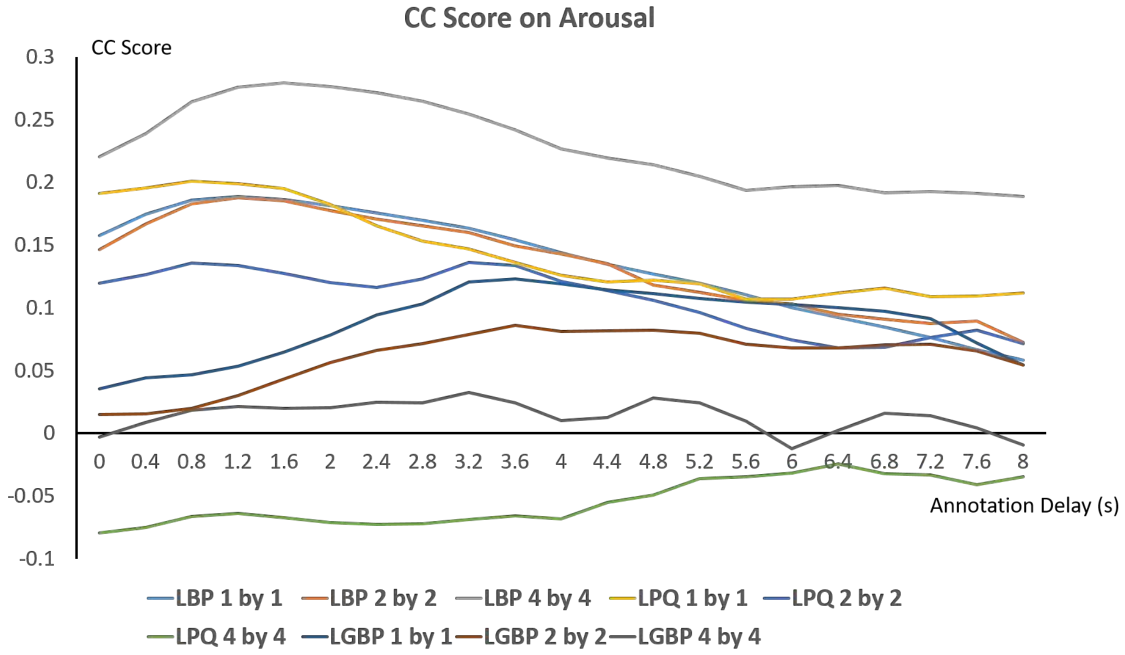


Figure 4.4: Plot of arousal delay for static features

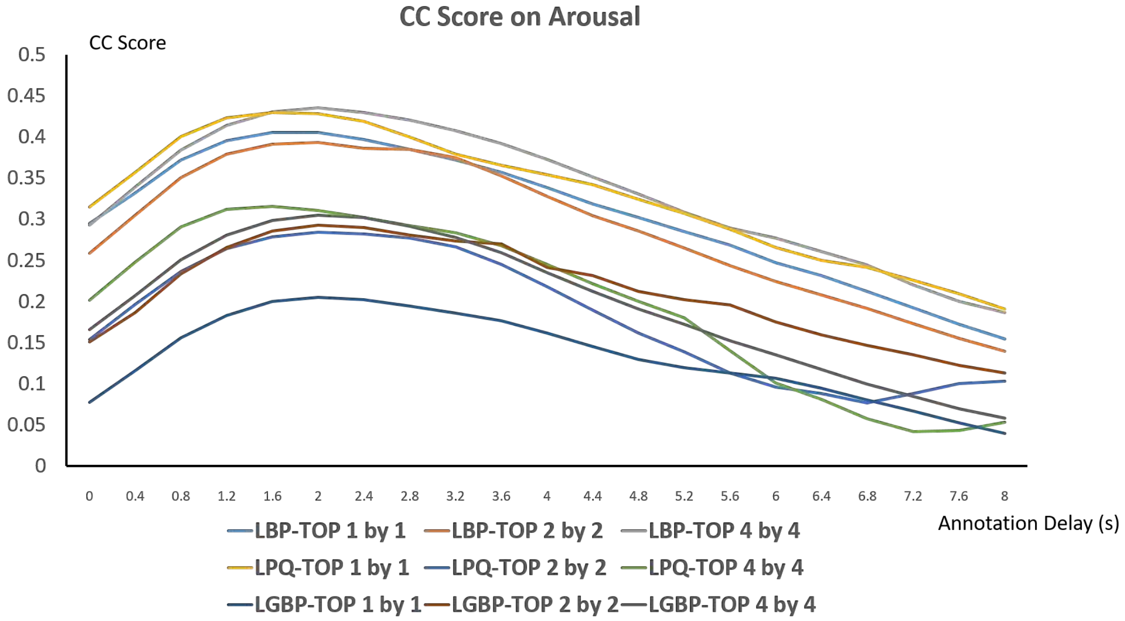


Figure 4.5: Plot of arousal delay for dynamic features

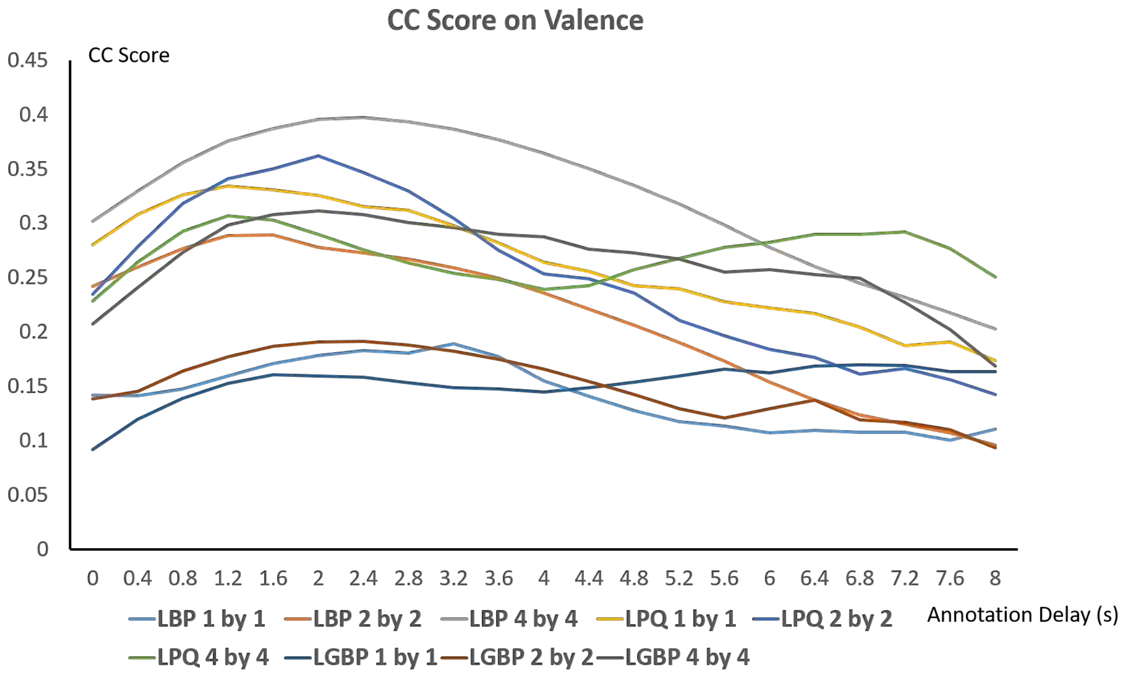


Figure 4.6: Plot of valence delay for static features

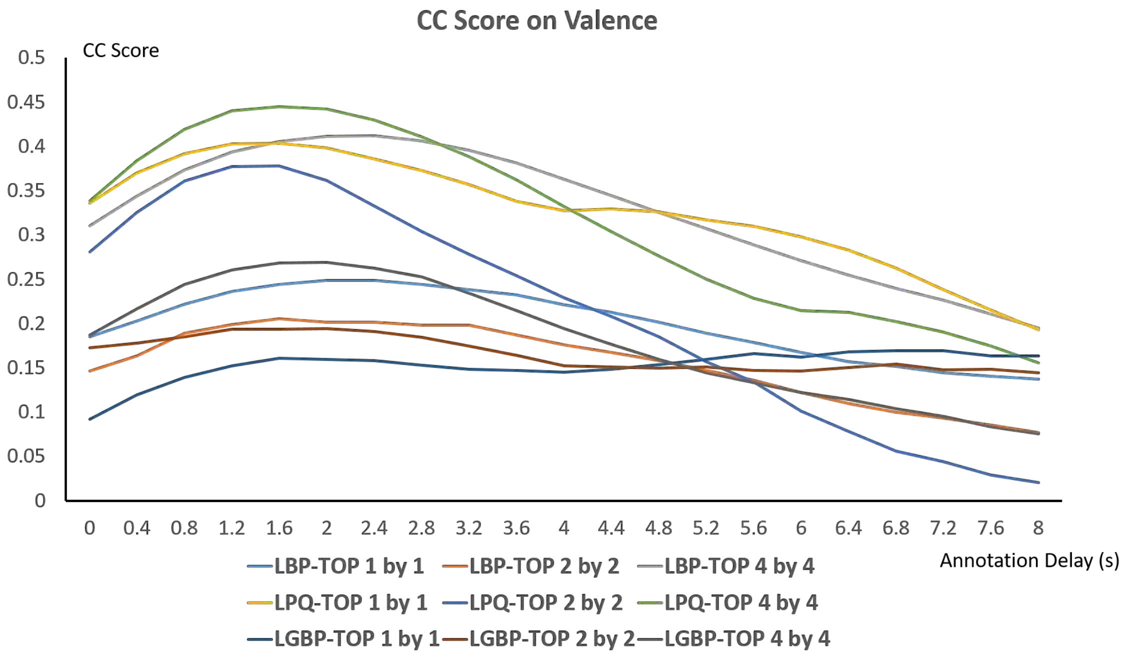


Figure 4.7: Plot of valence delay for dynamic features

4.4.1 Experimental Procedure

Similar to experiment, the AVEC 2015 dataset was used for this experiment. For this experiment, the following post-processing steps are applied to the initial predictions: (i) median filtering and (ii) centering and scaling. The median filter is a nonlinear filter which is commonly used in digital signal and image processing for noise reduction. The main idea of the median filter is to go through every entry in the signal and replace each entry with the median of its neighbouring entries. The size of neighbouring entries is usually referred as the window size. In this experiment, various window size have been tested ranging from 0.4s (10 frames) to 20s (500 frames) with a step size of 0.4s. For centring and scaling, the following formula was used.

$$y_{final} = \frac{(y_{pred} - \mu_{pred)}}{\sigma_{pred}} * \sigma_{train} + \mu_{train} \tag{4.1}$$

where y_{final} is the final prediction, y_{pred} is the raw prediction, mu_{pred} is the mean of the raw prediction, σ_{pred} is the standard deviation of the raw prediction, σ_{train} is the standard deviation of the training labels, and μ_{train} is the mean of the training labels.

The prediction results from the second experiment were used in this experiment, specifically the 4×4 LBP-TOP feature was selected for arousal prediction while the 4×4 LPQ-TOP feature was selected for valence prediction since they achieved the highest CC scores respectively after delay compensation.

4.4.2 Experimental Results and Analysis

Table 4.7 shows the original prediction results and the prediction results after post-processing. As it can be seen, both the arousal and valence prediction have been improved in term of CC score and largely improved in terms of CCC score. Through the experiment, it was found that for the arousal dimension, the median filter with window size of 4.8s achieved best results, while for the valence dimension, a window size of 4s achieved best results. To visualise the effect of the post-processing steps, Figure 4.8 and 4.9 show the arousal and valence curves of a segment of video, including the ground-truth labels, the initial prediction result, the prediction after median filter was applied and the prediction after centering and scaling. From the figures it can be seen, after the median filter, the prediction curve becomes much smoother while still capturing the trend of the ground truth. After the centering and scaling steps, the bias and scaling issues that come with the initial prediction have been reduced.

Table 4.7: Original and post processed result on the development partition of AVEC 2015 dataset

	Original			Post Processed		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP-TOP Arousal	0.184	0.435	0.225	0.181	0.590	0.553
LPQ-TOP Valence	0.111	0.445	0.303	0.125	0.546	0.537

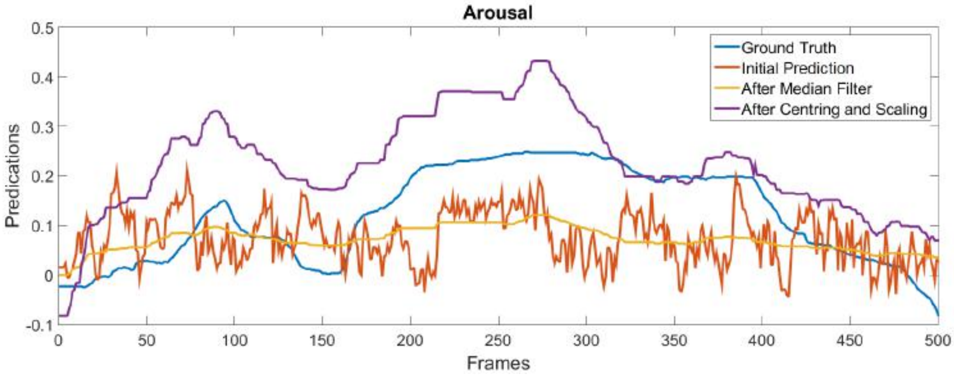


Figure 4.8: Plot of post processed prediction on arousal dimension using LBP-TOP features

Table 4.8 and 4.9 compare the recognition results with the AVEC 2015 baseline system and the AVEC 2015 winning system where the best results are highlighted in bold. It can be seen that our system outperformed the AVEC 2015 baseline system on both arousal and valence dimensions. Compared to the winning system in AVEC 2015, the proposed system achieved close performance on the arousal dimension and better performance on the valence dimension. However, it should be noted that the

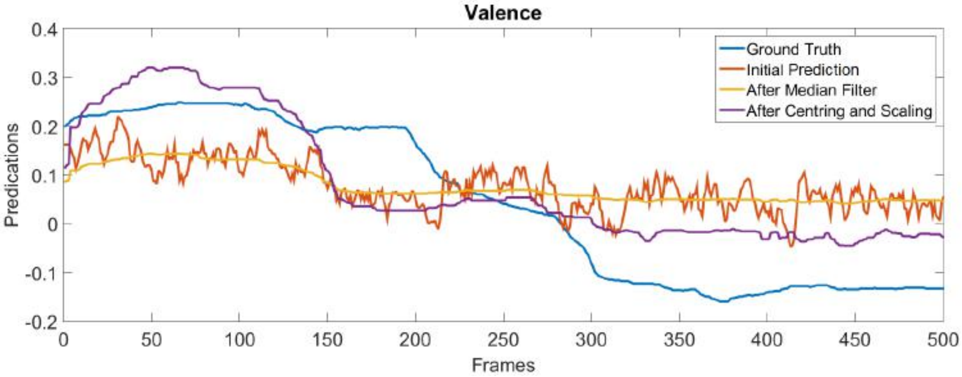


Figure 4.9: Plot of post processed prediction on valence dimension using LPQ-TOP features

Table 4.8: Comparison with selected baseline results on the development set on arousal dimension

	Arousal		
	Ringeval *et al.* (2015b)	He *et al.* (2015)	our results
RMSE	0.214	**0.148**	0.181
CC	0.183	**0.665**	0.590
CCC	0.103	**0.587**	0.553

winning system used a deep Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Network which is known to have a better performance on modelling temporal information compared to a linear SVR. It also used additional steps such as feature selection to select more correlated features. Since the aim of these experiments is to thoroughly investigate the performance of different histogram-based features on continuous affect recognition, only the linear SVR was selected as the learning technique in order to benchmark different features.

Table 4.9: Comparison with selected baseline results on the development set on valence dimension

	Valence		
	Ringeval *et al.* (2015b)	He *et al.* (2015)	our results
RMSE	0.117	**0.105**	0.125
CC	0.358	0.501	**0.546**
CCC	0.273	0.346	**0.537**

4.5 Experiment 4: Generalisation across datasets

In previous experiments it was investigated how block size effects the performance of different histogram-based features. The aim of this experiment is to investigate if a particular histogram-based feature will perform consistently well across different datasets. For AVEC 2015 dataset, the 4×4 LBP-TOP feature achieved the best result on arousal dimension, while the 4×4 LPQ-TOP achieved the best result on valence dimension after the delay compensation. In this experiment, the best block size configuration for different histogram-based features was selected based on previous experimental results. The selected configurations are: 4×4 LBP feature, 1×1 LPQ feature, 4×4 LGBP feature, 4×4 LBP-TOP feature, 1×1 LPQ-TOP feature, 4×4 LPQ-TOP feature and 4×4 LGBP-TOP feature. For this experiment, the aforementioned features were tested using the AVEC 2012 dataset and the DCU dataset introduced in Chapter 3.

4.5.1 Experimental Procedure

The DCU dataset was first partitioned into three person independent subsets (training, development and testing) similar to the AVEC 2012 and AVEC 2015 dataset. The ground truth annotation was calculated as the average across all three annotators. The same pre-processing steps used in Section 4.2 was first employed to extract and re-size the face region to 96×96. Next, different histogram-based features were extracted using the selected configurations and the same SVD from a low-rank approximation was used to reduce the feature dimensions. Finally, a linear SVM was used for prediction. To compare the performance with other baseline systems, the evaluation was carried out using the training partition and evaluated using the development partition for both the AVEC 2012 and DCU datasets.

Table 4.10: Recognition result for different histogram-based features on AVEC 2012 development partition

	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP (4×4)	0.235	0.250	0.153	0.251	0.377	0.308
LPQ (1×1)	0.238	0.242	0.173	0.241	0.205	0.157
LGBP (4×4)	0.226	0.355	**0.262**	0.274	0.294	0.262
LBP-TOP (4×4)	0.227	0.343	0.245	0.247	0.422	0.255
LPQ-TOP (1×1)	0.233	0.301	0.229	**0.238**	**0.453**	**0.411**
LPQ-TOP (4×4)	**0.222**	**0.403**	0.254	0.249	0.409	0.357
LGBP-TOP (4×4)	0.224	0.383	0.284	0.264	0.227	0.150

4.5.2 Experiment Results and Analysis

The recognition results for the AVEC 2012 and DCU datasets are shown in Table 4.10 and 4.11 respectively. For the AVEC 2012 dataset, it can be seen the 4×4 LPQ-TOP feature achieved best recognition result on arousal dimension, while the 1×1 LPQ-TOP feature performed best on the valence dimension in terms of CC score. For the DCU dataset, the 1×1 LPQ-TOP and 4×4 LBP-TOP feature achieved best recognition results on arousal and valence dimensions respectively in terms of both CC and CCC score. However, the recognition results on the valence dimension is relatively low compared to the arousal dimension on the DCU dataset. This could be caused by inconsistent annotations from multiple annotators on valence dimension as discussed in Section 3.5.

The recognition result for both datasets after delay compensation are shown in Table 4.12 and 4.13. Similar to the experiment without delay composition, 1×1 LPQ-TOP and 4×4 LBP-TOP achieved the best recognition results on arousal and valence dimensions respectively for the DCU dataset, while the 4×4 LPQ-TOP and

Table 4.11: Recognition result for different histogram features on DCU development partition

	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP (4×4)	0.277	-0.019	-0.016	0.208	-0.014	0.010
LPQ (1×1)	0.244	0.189	0.118	0.229	0.112	0.058
LGBP (4×4)	0.223	0.399	0.320	0.251	-0.122	-0.073
LBP-TOP (4×4)	0.217	0.426	0.303	**0.189**	**0.246**	**0.166**
LPQ-TOP (1×1)	**0.197**	**0.559**	**0.468**	0.237	0.171	0.093
LPQ-TOP (4×4)	0.268	0.470	0.262	0.240	-0.047	-0.026
LGBP-TOP (4×4)	0.203	0.522	0.409	0.223	0.065	0.039

the 1×1 LPQ-TOP features perform best for the AVEC 2012 dataset. In addition, it can be seen for the AVEC 2012 dataset that there exists a delay between 3 and 4 seconds for the arousal dimension and between 4 and 5 seconds for the valence dimension. For the DCU dataset, the delay is generally between 3 and 4 seconds for the arousal dimension while for the valence dimension the delay time is between 1 and 2 seconds.

Table 4.14 and Table 4.15 show the recognition results for both AVEC 2012 and DCU dataset after the post-processing steps proposed in Section 4.4 have been applied to the predictions after delay compensation. As can be seen, the post-processing steps have improved the performance of the proposed system on both datasets. Compared to the AVEC 2012 winning system, the proposed system has achieved better results on the valence dimension and comparable results on the arousal dimension in terms of CC score.

Table 4.12: Recognition result for different histogram features on AVEC 2012 development partition with delay compensation

	Arousal				Valence			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP (4×4)	2.4	0.232	0.266	0.166	5.6	0.248	0.419	0.359
LPQ (1×1)	1.6	0.236	0.255	0.184	4.4	0.263	0.291	0.226
LGBP (4×4)	2.4	0.222	0.385	0.297	3.6	0.271	0.331	0.305
LBP-TOP (4×4)	4	0.221	0.379	0.277	4	0.245	0.449	0.282
LPQ-TOP (1×1)	4.8	0.228	0.325	0.253	4.8	**0.235**	**0.493**	**0.437**
LPQ-TOP (4×4)	4	**0.217**	**0.426**	0.280	4	0.245	0.444	0.395
LGBP-TOP (4×4)	3.6	0.219	0.409	**0.299**	3.6	0.263	0.251	0.174

Table 4.13: Recognition result for different histogram features on DCU development partition with delay compensation

	Arousal				Valence			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP (4×4)	0	0.277	-0.019	-0.016	2.8	0.207	0.013	0.009
LPQ (1×1)	4	0.261	0.196	0.105	1.2	0.230	0.112	0.058
LGBP (4×4)	2	0.218	0.429	0.350	0.8	0.250	-0.120	-0.072
LBP-TOP (4×4)	2.4	0.212	0.474	0.362	1.6	**0.189**	**0.258**	**0.178**
LPQ-TOP (1×1)	3.2	**0.182**	**0.644**	**0.572**	2	0.237	0.181	0.094
LPQ-TOP (4×4)	3.2	0.266	0.528	0.302	1.2	0.241	-0.041	-0.022
LGBP-TOP (4×4)	3.2	0.198	0.563	0.462	0.8	0.223	0.066	0.040

Table 4.14: Recognition result on AVEC 2012 development partition after post-processing

	Arousal			**Valence**		
	Before	After	Nicolaou *et al.* (2012)	Before	After	Nicolaou *et al.* (2012)
RMSE	0.217	0.227	-	0.235	0.228	-
CC	0.426	0.516	0.538	0.493	0.539	0.354
CCC	0.280	0.512	-	0.437	0.488	-

Table 4.15: Recognition result on DCU development partition after post-processing

	Arousal		**Valence**	
	Before	After	Before	After
RMSE	0.182	0.175	0.189	0.246
CC	0.644	0.777	0.258	0.326
CCC	0.572	0.725	0.178	0.302

4.6 Experiment 5: Affect Recognition using Convolutional Neural Network

As discussed in Section 2.4.2.2, a number of well-established pattern recognition problems such as object detection, image recognition and speech recognition have benefited greatly from the advent of deep neural networks. Recently, deep neural networks have also been applied to continuous affect recognition and have achieved state-of-the-art performance. The deep neural networks are usually used in three different ways. The first uses the deep neural network as a feature extractor. For instance, Chao *et al.* (2015) trained the Convolutional Neural Networks (CNN) on 110,000 images from 1032 people in Celebrity Faces in the Wild (CFW) (Zhang *et al.*, 2012) and FaceScrub (Ng and Winkler, 2014) datasets. The 9216 nodes' values from the last convolutional layer are then used to compute face features and are used to train a Long Short Term Memory Recurrent Neural Network (LSTM) for affect recognition. The second uses the deep neural networks as a machine learning method. The work carried out by He *et al.* (2015) uses LGBP-TOP features extracted from the face region and LSTM to predict the affective state. The third approach treats the deep neural network as a combination of feature extraction and machine learning. For example, Khorrami *et al.* (2016) uses the facial images directly from the dataset to train the CNN for affect recognition.

However, to the best of our knowledge the current deep learning approaches all use features extracted from the entire face region, and none of them have considered to use individual facial parts such as eyes and mouth for affect recognition. People tend to use different facial parts as reference when perceiving other people's emotions. For instance, the work carried out by Jack *et al.* (2012) suggested that the mouth is more informative for western people when judging people's affective state whereas the eyes are more informative for East Asian people. The aim of this experiment is to investigate if it is possible to use individual facial parts for continuous affect recognition. In this experiment, the CNNs are used as a combination of feature extraction and machine learning. They are trained using both the entire face and individual facial parts as input and their respective performances are compared.

4.6.1 Data Pre-processing

The AVEC 2015 dataset was chosen for this experiment. In total, 67,500 images are used for training, 30,000 images are used for validation and 47,500 images are used for testing. For CNN training using the entire face image, the face region was extracted using the same method introduced in Section 4.2.1 while the face image was also re-sized to 96×96. For the CNN trained using facial parts, each facial part was extracted using the detected landmarks. In particular, the following facial parts were selected for this experiment: (i) left and right eye region to capture gaze direction, blinks and eyebrows movements, (ii) glabellar lines to capture extreme negative affective state such as anger, fear and depression, (iii) left and right nasolabial folds to capture smiles, and (iv) mouth to capture mouth movements. Each of these facial parts are shown in Figure 4.10. Before training, the input images for both networks

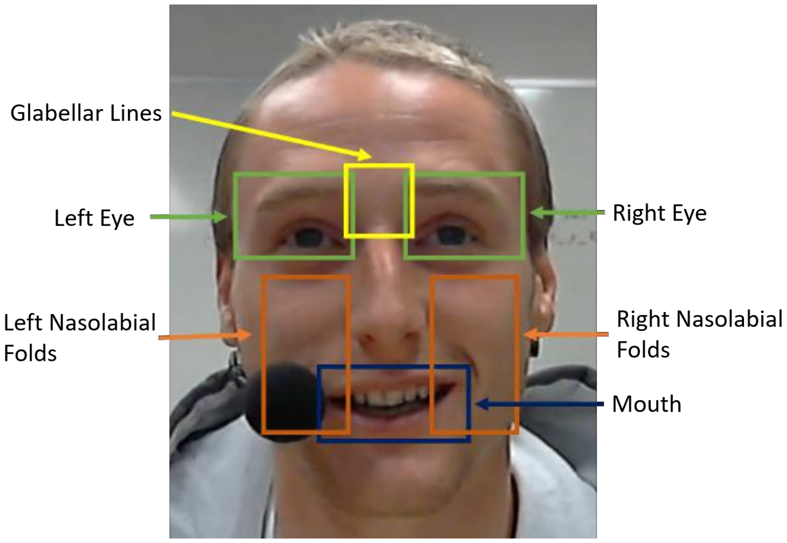


Figure 4.10: Facial Part

are standardised by subtracting the mean and dividing by the standard deviation computed on the training partition. All experiments carried out in this section were trained using an NVIDIA GTX 970 GPU with 4 Gigabyte memory.

4.6.2 Experimental Procedure

A classic feed-forward CNN was used in all experiments presented in this section. For simplicity, the CNN trained using an entire face is referred as CNN_{entire} while the CNN trained using facial parts is referred to as CNN_{part}. The network structure for CNN_{entire} is similar to one used by Khorrami *et al.* (2016) since it's one of the first works using CNN for End-to-End affect recognition, and achieved better results when compared to hand-crafted features (Ringeval *et al.*, 2015a,b). The network consists of three convolution layers with 64, 128, 256 filters respectively, with filter sizes of 5×5. Each layer is followed by a max pooling layer of size 2×2. The network is then followed by a fully-connected layer with 300 hidden units and a dense layer with output equal to 1. CNN_{part} consists of four convolution layers with 64, 64, 128, 256 filters respectively, with filter sizes of 3×3. A max pooling layer with size 2×2 is applied after each convolution layer except for the first one. The network is then followed by a fully-connected layer with 128 hidden units and, finally, a dense layer with output equal to 1 was used to predict either the arousal or the valence value. The network was trained for each facial part separately. For all hidden layers in CNN_{entire} and CNN_{part}, the ReLU activation function was used. Both CNN_{entire} and CNN_{part} were trained using stochastic gradient descent with a batch size of 128, momentum of 0.9 and a weight decay of 1e-5. The learning rate was set to 0.001 throughout the training. The parameters of each layer were initialised with uniform distribution.

As shown in Section 4.3, there exist delays in the annotations for the AVEC 2015 dataset and the experiments with histogram-based features have shown that the recognition results were significantly improved after delay compensation. In this experiment the same delay compensation method was used by shifting the annotation forward from [0-8] seconds with a step size of 0.4 seconds. The delayed annotations were then used to train both CNN_{entire} and CNN_{part}.

4.6.3 Experimental Results and Analysis

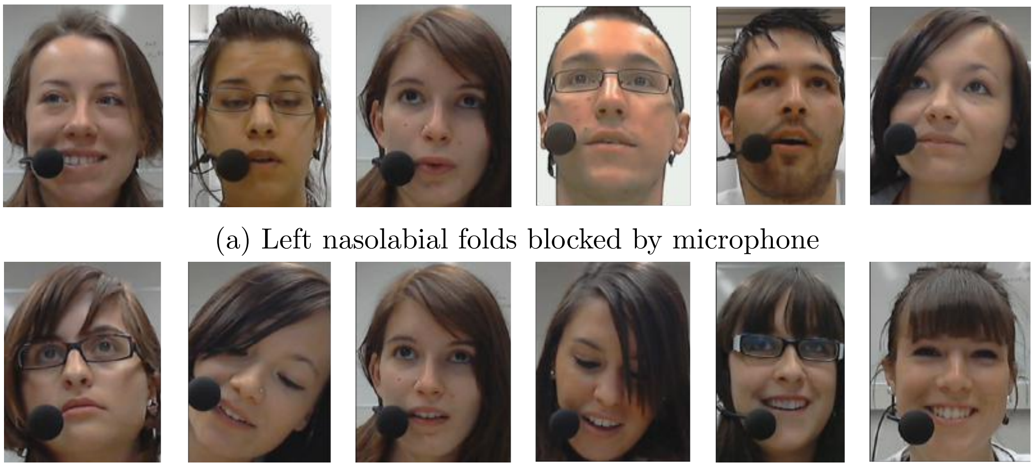
Table 4.16 shows the performance of the CNN_{entire} and CNN_{part} in terms of CC score. The results shown that not only is it possible to use facial parts for continuous affect recognition, but also this approach achieved better performance compared to the entire face approach. It can be seen from the results when using facial parts that the mouth region is good at predicting the valence dimension, while the eye regions and glabellar lines achieved better results on the arousal dimension. The performance of nasolabial folds are similar for both arousal and valence dimensions. Although it is expected that the glabellar lines should perform well on the valence dimension since it is highly correlated with negative emotions, the recognition results show poor performance on the valence dimension. This might be caused by the fact that the training data only consists of very few examples of negative emotions which involve glabellar lines, meaning the network is unable to learn any useful features. When using the entire face, the performance on the valence dimension is significantly better than the arousal dimension. Among different facial parts, the right nasolabial folds achieved best results on the arousal dimension, while the mouth performs best for the valence prediction.

Since a face is symmetric, one might expect that the performance of symmetric parts should be similar. However, the results have shown that the right nasolabial folds perform better compared to their symmetric part. After analysing the data it is found that the main reason for this is due to the visibility of the facial part. For instance, the left nasolabial region is constantly obstructed by the microphone worn by the participants as shown in Figure 4.11a. The result is that the performance of left nasolabial folds is significantly lower compared to the right one. Since both left eye and right eye can be blocked by hair (See Figure 4.11b), they achieved very similar results.

Table 4.16: CC score without annotation delay using CNN

	Arousal	Valence
Whole Face	0.014	0.243
Mouth	0.032	**0.334**
Left Eye	0.122	0.017
Right Eye	0.177	0.060
Glabellar Lines	0.154	-0.083
Left Nasolabial Folds	-0.080	-0.023
Right Nasolabial Folds	**0.198**	0.132

Table 4.17 shows the recognition results in terms of CC score after delay compensation has been applied. Similar to previous experiments, the performance for both CNN_{entire} and CNN_{part} have been improved. The eye regions and mouth region are generally good at predicting both arousal and valence dimensions compared to other facial parts. This agrees with the findings by Jack *et al.* (2012) who suggest that both mouth and eyes are informative in judging people's emotions. After delay compensation, the right eye achieved best results on the arousal dimension



(a) Left nasolabial folds blocked by microphone

(b) Eye regions blocked by hair

Figure 4.11: Sample data from AVEC 2015 dataset

followed by the right nasolabial folds, while the mouth region achieved best results on the valence dimension. In addition, it can be seen that the different facial parts have different time delays when predicting arousal and valence dimensions. This might suggest that people have different reaction time when perceiving emotions from different facial parts. The mouth region requires a longer reaction time when judging the arousal dimension compared to the valence dimension. In contrast the nasolabial folds requires longer reaction time on the valence dimension than arousal. Compared to other facial parts, the eye regions have shorter reaction time for both dimensions.

Table 4.17: CC score with annotation delay using CNN

	Delay Time	Arousal	Delay Time	Valence
Whole Face	0.8	0.114	1.2	0.401
Mouth	2.4	0.258	0.8	**0.432**
Left Eye	0.8	0.250	0.8	0.233
Right Eye	0.8	**0.274**	0.8	0.157
Glabellar Lines	2.4	0.223	0	0.016
Left Nasolabial Folds	0.8	0.031	1.6	0.104
Right Nasolabial Folds	0.8	0.244	2.4	0.204

Compared to the previous experiments that use histogram-based features, the use of facial parts and CNN have shown comparable results on the valence dimension

Table 4.18: Results comparison between histogram-based features and facial parts in terms of CC score

Arousal		**Valence**	
LBP-TOP	Right Eye	LPQ-TOP	Mouth
0.435	0.274	0.445	0.432

but lower results on the arousal dimension (See Table 4.18). This might due to the fact that in order for a CNN to learn representative features, a very large amount of data with significant variations is usually needed.

4.7 Conclusion

In this chapter various experiments have been carried out to investigate thoroughly the performance of different histogram-based features in continuous affect recognition and to study if individual facial parts can be used for continuous affect recognition.

The experimental results indicate that unlike facial action recognition where LGBP-TOP features outperformed other histogram-based features, the LBP-TOP and LPQ-TOP features in general perform best for continuous affect recognition. However, the best block size configurations tend to vary for different datasets and dimensions when measured in terms of CC score. For the AVEC 2015 dataset, the 4×4 LBP-TOP features performs best for the arousal dimension and the 4×4 LPQ-TOP features perform best for valence dimension. For the AVEC 2012 dataset, the 4×4 LPQ-TOP features achieved best results on arousal dimension while the 1×1 LPQ-TOP features perform best for the valence dimension. For the DCU dataset, the 1×1 LPQ-TOP features obtained the highest CC score on the arousal dimension while the 4×4 LBP-TOP features perform best for valence dimension. By combining the best histogram features with Support Vector Regression (SVR) and post-processing techniques, the proposed system has achieved better results on the valence dimension and comparable result on the arousal dimension in terms of both CC score when compared to the AVEC 2012 and AVEC 2015 winning systems.

The annotation delay analysis shows that the annotation delay also tends to vary for different datasets and dimensions. In general, the delay time is longer for the valence dimension than the arousal dimension. It was observed for the AVEC 2015 dataset that the delay time are between 1.6 and 2 seconds on the arousal dimension and 1.6 to 2.4 seconds on the valence dimension while for the AVEC 2012 dataset the delay time are between 3 and 4 seconds on the arousal dimension and 4 to 5 seconds on the valence dimension. For the DCU dataset, the delay time are between 2.4 to 2.8 seconds for the arousal dimension and 1 to 2 seconds for the valence dimension.

The experiments on individual facial parts have shown that by using Convolutional Neural Networks (CNN) it is not only possible to predict affective state using facial parts, but also that it achieved better results when compared to the use of the entire face image directly as input. In general, the eye region is good at predicting arousal while the mouth region performs best for valence prediction. In addition, this approach should be more robust to occlusion compared the use of the entire face since one could use the facial parts that are not occluded for affect prediction. The annotation delay analysis suggests that when annotating the data the reaction time for different dimensions is different. The mouth region requires a longer reaction time when judging arousal than valence and the nasolabial folds require longer reaction time for valence than arousal. Compared to histogram-based features, the use of facial parts achieved comparable results on the valence dimension but lower results on the arousal dimension which might due to the fact that the amount of training data for the CNN to learn representative features was relatively small.

Chapter 5

Multi-Modal Affect Recognition

5.1 Introduction

As suggested by Ambady and Rosenthal (1992), facial expression gives the most clear and naturally preeminent signals for humans to communicate emotions. It is used for clarification, giving emphasis, expressing intentions and more generally, to regulate interactions under different environments and other people. These facts highlight the importance of facial behaviour analysis in automatic affect recognition. As discussed in Section 2.2 and Section 2.3.4, much of the previous research has focused on recognising deliberately displayed affective state, mainly prototypical expression of six basic emotions, captured under highly controlled environments. Recent efforts focus on the recognition of complex and spontaneous affective state which is annotated continuously using dimensional models. When using visual modalities, most of these existing systems use 2D facial images which usually require to maintain a consistent frontal face view in order to achieve good recognition performance. In addition, systems that use 2D facial images are also sensitive to recording conditions such as illumination and occlusions. One advantage of using a depth image is its robustness against different lighting conditions as shown in Figure 5.1. It can be seen that under low light conditions, it is impossible to detect the face from the colour image, however, the depth image is not affected by the lighting condition. Furthermore, since a 2D facial image is unable to capture out-of-plane changes, certain facial expressions such as lip pucker and jaw clenching are difficult to detect in a 2D frontal view (Sandbach *et al.*, 2012). To tackle these problems, depth (3D) data can be used. Although, various 3D affect datasets have been captured, there still does not exist any publicly available 3D affect dataset that includes recordings of spontaneous behaviours that is also annotated continuously using dimensional models. As a result, the reported research using depth data only focuses on discrete affect recognition.

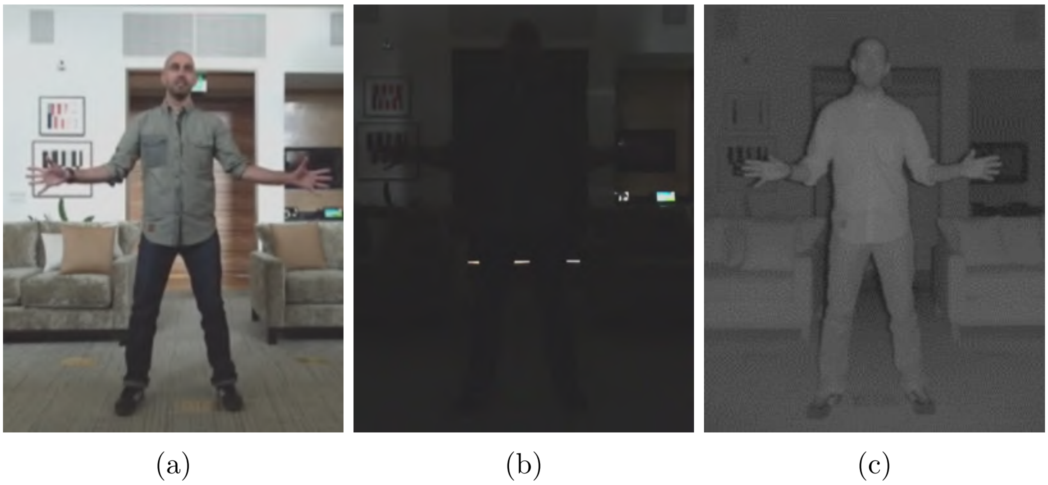

(a) (b) (c)

Figure 5.1: Comparison between colour and depth image under different lighting conditions. (a) and (b) shows the colour image with light on and off. (c) shows the depth image with light off

In addition, recent works on continuous affect recognition have shifted towards utilising multiple modalities. For instance, the work carried out by Valstar and Pantic (2012) and Nicolle *et al.* (2012) investigated the use of video and audio modalities, while the work by Ringeval *et al.* (2015b) and He *et al.* (2015) studied the use of additional physiological modalities. It has been shown that by fusing different modalities, the recognition results are generally improved when compared to using a single modality. As discussed in Section 2.4.3, two of the most commonly used fusion methods are feature-level fusion and decision-level fusion. In automatic affect recognition, feature-level fusion is obtained by concatenating all the features from different modalities into one feature vector which is then fed into machine learning techniques. When the frame rate from different modalities are different, down-sampling or up-sampling is used to ensure all modalities have the same frame rate. In decision-level fusion, the input for each modality is modeled independently, and the single modal recognition results are combined in the end, using machine learning techniques.

By using the dataset introduced in Chapter 3, the aim of this Chapter is to first investigate how 3D data can be used for continuous affect recognition and then to study if the use of video and depth modalities can further improve recognition results.

In order to use 3D facial data, the location of the face must be first detected. Section 5.2 proposes a method for face detection using the depth image that is then used to extract all the face regions from the recordings. In Section 5.3, the histogram-based features used in previous experiments are applied to the depth data, and the performance of these features are examined and compared with the results from the video modality. Section 5.4 proposes a multi-modal affect recognition framework which can be used in real world settings (e.g. under low light conditions). Experiments are then carried out to investigate if the use of both colour and depth features could lead to a significant increase in performance in affect recognition. In particular, both aforementioned fusion methods are evaluated.

5.2 Face Detection using Depth Data

As discussed earlier, previous 3D facial expression studies have been mainly carried out on publicly available 3D expression datasets such as BU-3DFE (Yin *et al.*, 2006) and BU-4DFE (Yin *et al.*, 2008) for discrete expression classification and recognition of facial action units (AUs). These datasets only capture the face region which means they can be used directly without the need to extract the face region. Unlike the aforementioned datasets, the dataset introduced in Chapter 3 captures the full scene instead of the face region, thus it is necessary to locate the face in order to use it for continuous affect recognition. One way to do this is by aligning the depth image with the colour (2D) image using camera calibration. After the depth image is aligned, the face detection result on the colour image can be projected on to the depth image to locate the face. However, this approach is limited by the fact that face detection in colour images is highly sensitive to illumination conditions which means under low light conditions, the face detection may no longer work. Compared to colour images, depth images are more robust to illumination changes which means the depth data can be used when the 2D facial image is not visible. In order to leverage this it is necessary to detect the face location directly on the depth image.

Various methods have been proposed for face detection using depth data. For instance, Colombo *et al.* (2006) performed 3-D face detection by first identifying candidate eyes and noses using curvature analysis, and then by using the candidate regions in a PCA-based classifier. In the work carried out by Mian *et al.* (2007), face detection is achieved by first finding the location of the nose tip, and then the face region is localised by a cropping sphere centred at the noise tip. Nair and Cavallaro (2009) proposed using a point distribution model for face detection. Although different methods have been proposed, these methods usually require high resolution depth data which is different from the data provided by the Microsoft Kinect. In this section, a method is proposed to use the Histogram Of Gradient features (Dalal and Triggs, 2005) combined with a structural SVM based training algorithm King (2015) to locate faces in the low resolution depth image obtained from a Microsoft Kinect.

5.2.1 Data Collection and Annotation

In order to train the face detector, 420 depth images with various head poses are extracted from the dataset captured in Chapter 3. On average, 30 depth images are extracted for each participant. The 420 depth images are then annotated manually by drawing a bounding box around the face region. The images are then split into person independent groups as shown in Figure 5.2 where each group consists of 30 images of the same participant.

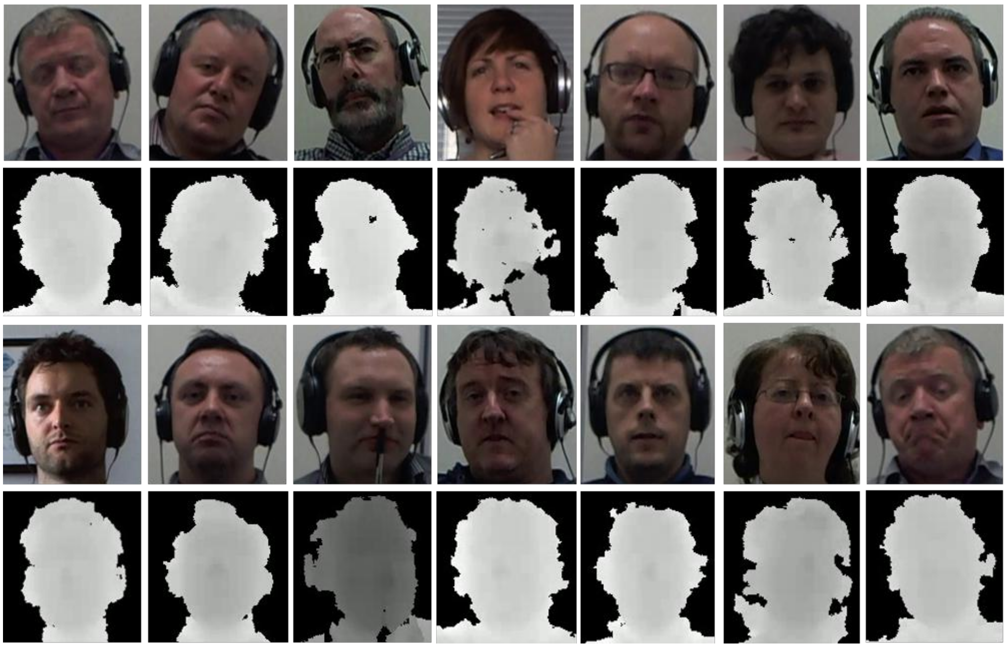


Figure 5.2: Samples of extracted depth image and the corresponding colour image.

5.2.2 Data Pre-processing

To increase the detection accuracy, the openNI library was first used to remove the background from the depth image. This is shown in Figure 5.3. Before training, each image is up-sampled by a factor of 2 to allow detection of small faces, followed by adding a mirrored version of each training image since human faces are generally left-right symmetric, thus doubling the number of images to 840. The range of the raw depth data is from 0 to 4096, it is then normalised to the range of 0 to 255 (8 bit).

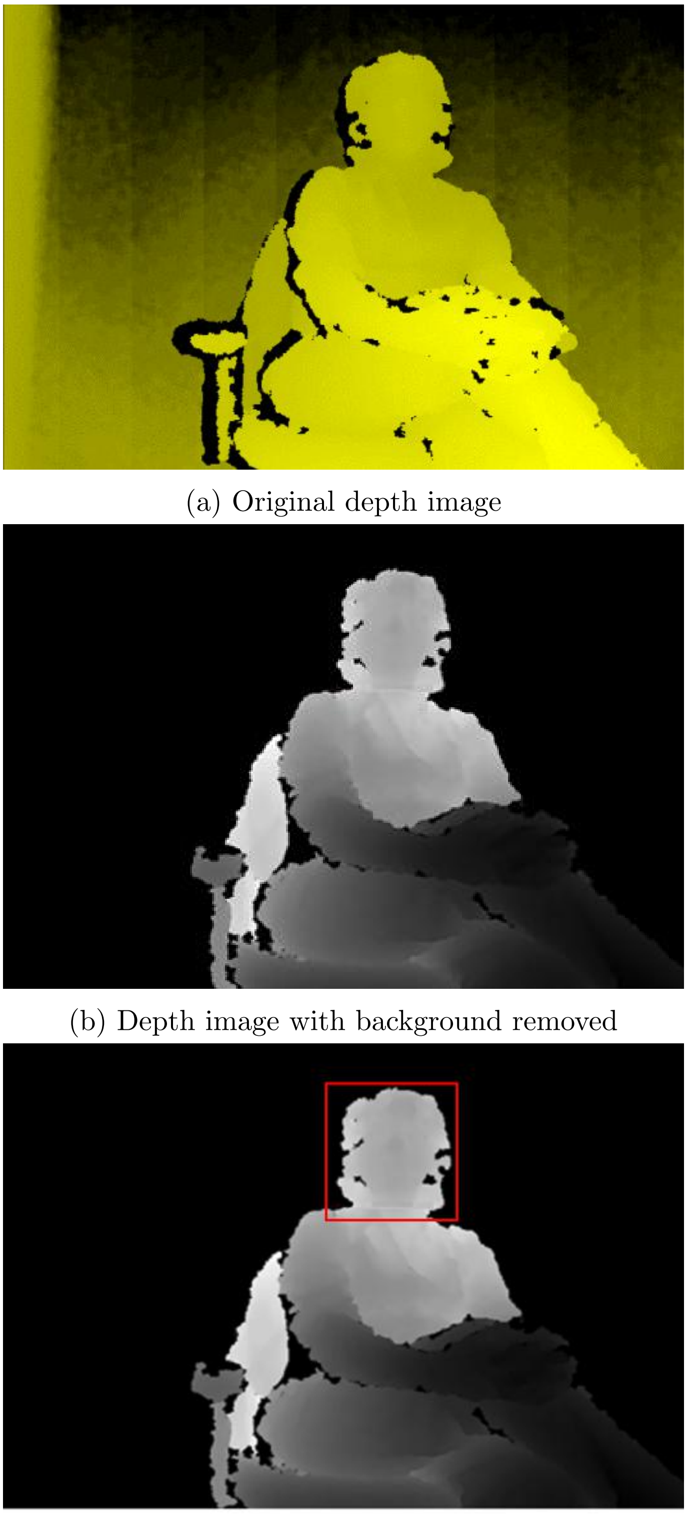
5.2.3 Experimental Procedure

To extract the HOG features, an image pyramid that down-samples the image at a ratio of 5/6 was applied to each image. For each pyramid level a sliding window with size 80×80 is applied to each image and the HOG features were extracted. The structural SVM based training algorithm (King, 2015) was used to train the face detector. The complexity parameter (C) was set to 1 and the epsilon was set to 0.01. The Dlib library (King, 2009) was used throughout this experiment for both HOG feature extraction and SVM learning. The 5-fold cross-validation leave-one-out method was used to compute the accuracy of the face detector. In order to test the generalisability of the face detector, cross-validation was applied to person independent groups instead of all extracted images, which means the training set and testing set do not contain the same person.

5.2.4 Experimental Results and Analysis

Figure 5.4 shows a visualisation of the learnt HOG descriptors. Due to noise and the limited accuracy of the Kinect sensor, both 16 bit and 8 bit depth images give good face boundary details though they provide less detail around the centre part of the face.

Table 5.1 shows the face detection results using different image types. Due to the relatively small number of testing subjects and the simple scene, all image types achieved very high face detection accuracy. Although the 16 bit depth image could identify more face structures compared to the 8 bit one since it has a bigger range, this did not improve the detection accuracy. Figure 5.5 shows some examples of the detection results, and it can be seen that the face detector has shown good



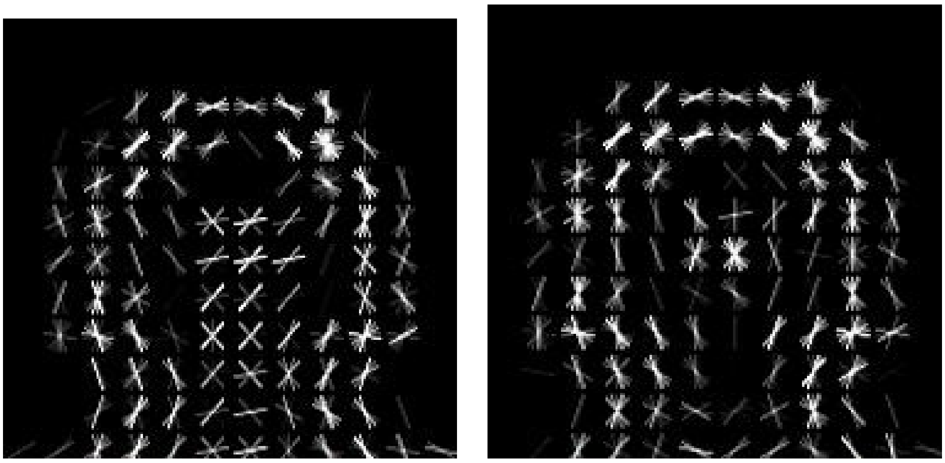
(a) Original depth image

(b) Depth image with background removed

(c) Labeled face on depth image

Figure 5.3: Pre processing steps

performance on various head poses.



(a) 16 Bit Depth

(b) 8 Bit Depth

Figure 5.4: Visualisation of learned HOG detector

Table 5.1: Face detection results using depth images

	16 Bit Depth	8 Bit Depth
Accuracy	97%	97%

5.3 Affect Recognition using Depth Data

There has been extensive research on solving the problem of 3D Facial Expression Recognition. As indicated by Sandbach *et al.* (2012), most of the systems developed have attempted to recognise expressions from static 3D facial expression data, while more recent works use dynamic 3D facial expression data. Various static and dynamic 3D features have been developed, as discussed in Section 2.4.1.2. However, previous systems have been only focused on predicting discrete affective state, and to the best of our knowledge, no research has been done on using the depth data for continuous affect recognition. The aim of this section is to investigate how depth data could be used for continuous affect recognition. Specifically, the histogram-based features used in Chapter 4 are applied to the depth data and the SVR is used to evaluate the performance of the features.

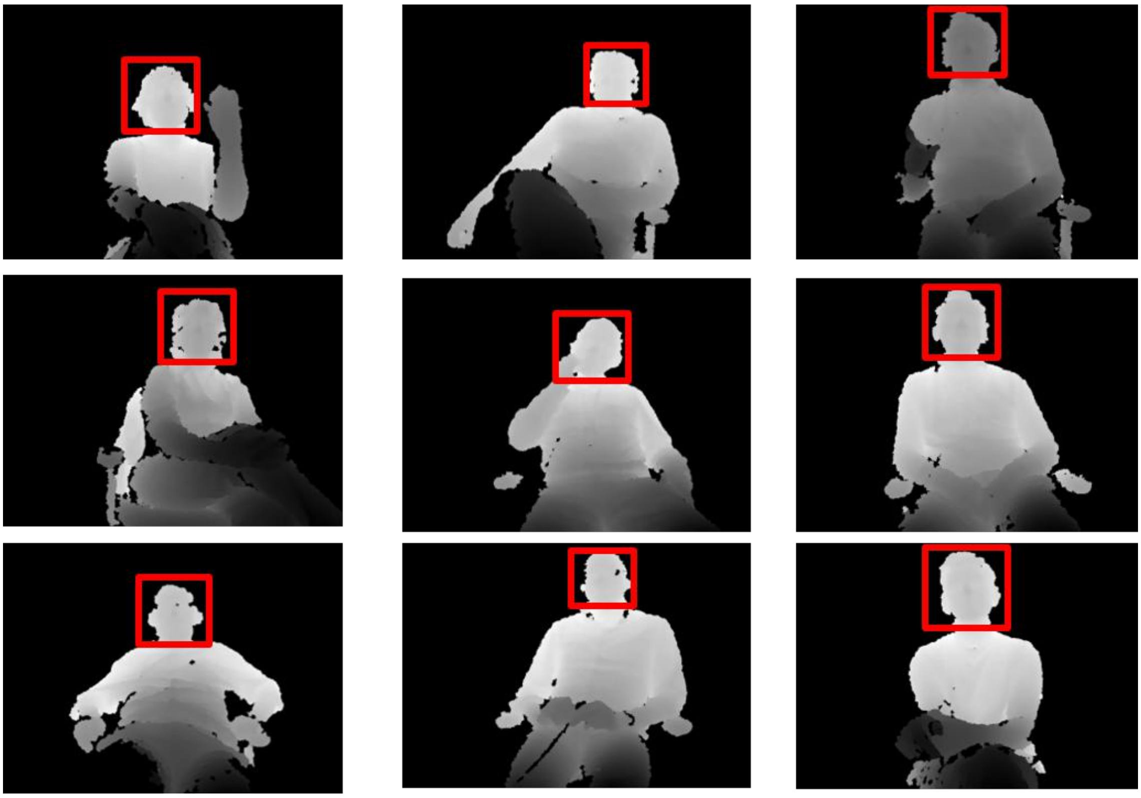


Figure 5.5: Example detection results using 8 bit depth image

5.3.1 Data Pre-processing

To extract the histogram-based features, the face detector proposed in Section 5.2 is first used to extract the face image. The face image is then re-sized to 96 × 96 for feature extraction. 2D Gaussian smoothing with a kernel size 5 × 5 is then applied to reduce the noise of the depth image. The overview of the pre-processing steps is shown in Figure 5.6.

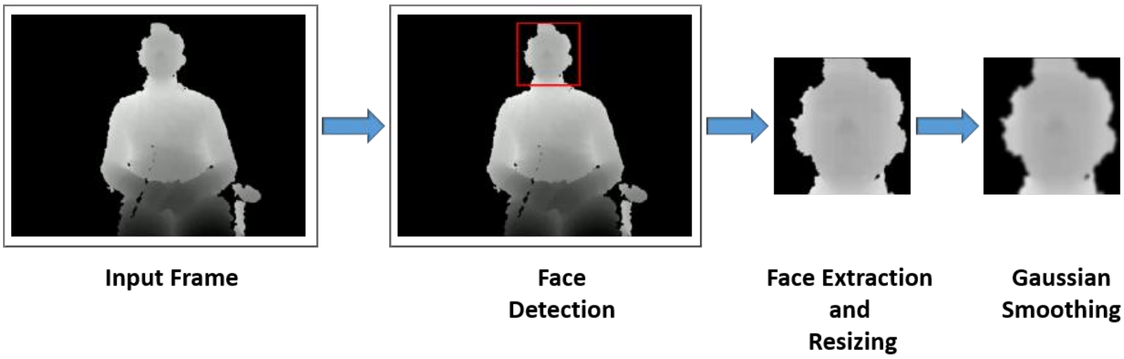


Figure 5.6: Overview of pre-processing steps

5.3.2 Experimental Procedure

Similar to previous experiments (Section 4.2), the LBP, LPQ, LGBP, LBP-TOP, LPQ-TOP and LGBP-TOP features with block size of 1 × 1, 2 × 2 and 4 × 4 are used in this experiment. The same configurations as shown in Table 4.2 are used to extract the histogram-based features. A SVD from a low-rank approximation is then applied to reduce the dimensions of the features. The original number of features and number of reduced features for each feature are shown in Table 5.2.

For the learning method, the SVR with the same parameter configurations as in Section 4.2 are used. The recognition results are measured in terms of Root Mean Square Error (RMSE: lower is better), Correlation Coefficient (CC: higher is better) and Concordance Correlation Coefficient (CCC: higher is better).

Table 5.2: Original number of features and reduced number of features for different block sizes

	1×1		2×2		4×4	
	Total	Reduced	Total	Reduced	Total	Reduced
LBP	59	19	236	17	944	19
LPQ	256	117	1024	78	4096	74
LGBP	177	79	1416	108	5664	112
LBP-TOP	177	71	708	54	2832	56
LPQ-TOP	708	296	3072	230	12288	229
LGBP-TOP	1416	442	2832	213	8496	150

5.3.3 Experiment Results and Analysis

As can be seen from Table 5.3, the 4×4 LPQ-TOP feature achieved best results in predicting the arousal and valence dimensions compared to other features. Similar to the experimental results when using video data, the depth data achieved good results on the arousal dimension, but performed poorly on the valence dimension which might be caused by the unreliable annotation on the valence dimension as discussed in Section 4.5.2.

Table 5.4 shows the recognition results after delay compensation has been applied. Among all features, the 4×4 LPQ-TOP achieved best results on both arousal and valence dimensions. The results suggest that the delay is between 0.4 to 2.4 seconds for the arousal dimension which is similar to the one obtained using the video modality. However, for the valence dimension, the delay tends to vary largely depending on which feature is used. This could be caused by unreliable annotation on the valence dimension as discussed earlier.

Table 5.5 shows the results after post-processing steps have been applied. It can be seen, as in previous experiments, that the post-processing steps have improved the recognition results in terms of both CC and CCC scores. However, the improvement is relatively small for the valence dimension.

Compared to the previous experimental results using the video modality, the results from the depth modality are relatively poor, which might be caused by the

Table 5.3: Results on development partition for different block sizes. The best results in different metrics are highlighted in bold

	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP	0.256	0.190	0.127	0.255	-0.202	-0.136
LPQ	**0.232**	**0.298**	0.227	0.245	-0.082	-0.050
LGBP	0.241	0.184	0.134	0.245	**0.099**	**0.052**
LBP-TOP	0.261	0.158	0.106	0.245	-0.175	-0.124
LPQ-TOP	0.234	0.294	**0.236**	0.261	-0.061	-0.038
LGBP-TOP	0.237	0.252	0.191	**0.232**	0.028	0.016

(a) Block size: 1 × 1

	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP	0.233	0.232	0.121	0.239	-0.010	-0.006
LPQ	**0.229**	0.301	0.196	**0.227**	0.078	0.046
LGBP	0.237	0.221	0.158	0.252	0.005	0.003
LBP-TOP	0.245	0.223	0.121	0.230	0.053	0.034
LPQ-TOP	0.231	**0.318**	0.198	0.245	**0.098**	**0.062**
LGBP-TOP	0.232	0.303	**0.236**	0.233	0.055	0.030

(b) Block size: 2 × 2

	Arousal			**Valence**		
	RMSE	CC	CCC	RMSE	CC	CCC
LBP	**0.225**	0.347	0.189	0.221	-0.040	0.024
LPQ	0.226	0.341	0.268	0.222	0.035	0.023
LGBP	0.232	0.296	0.228	0.247	0.022	0.010
LBP-TOP	**0.225**	0.353	0.194	**0.220**	-0.067	-0.040
LPQ-TOP	0.228	**0.361**	**0.305**	0.229	**0.112**	**0.071**
LGBP-TOP	0.235	0.278	0.227	0.236	0.054	0.031

(c) Block size: 4 × 4

Table 5.4: Results on development partition for different block sizes with shifted label

	Arousal				**Valence**			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP	0	0.256	0.190	0.061	2	0.255	-0.200	-0.135
LPQ	2.4	**0.232**	0.297	0.224	1.2	0.245	-0.079	-0.049
LGBP	4.8	0.240	0.196	0.139	4.4	0.246	**0.109**	**0.057**
LBP-TOP	0	0.261	0.158	0.106	2.4	0.247	-0.174	-0.124
LPQ-TOP	2.8	0.234	**0.305**	**0.243**	1.2	0.261	-0.060	-0.037
LGBP-TOP	1.6	0.236	0.261	0.199	3.2	**0.232**	0.031	0.018

(a) Block size: 1 × 1

	Arousal				**Valence**			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP	0.8	0.233	0.233	0.121	0	0.239	-0.010	-0.006
LPQ	0.8	**0.200**	0.136	0.090	2	**0.124**	**0.036**	**0.033**
LGBP	4.8	0.224	0.082	0.039	2.4	0.131	0.191	0.065
LBP-TOP	1.2	0.245	0.224	0.121	0	0.230	0.052	0.034
LPQ-TOP	2	0.230	**0.327**	0.209	0.4	0.245	0.098	0.062
LGBP-TOP	2	0.232	0.307	**0.233**	0.4	0.233	0.056	0.031

(b) Block size: 2 × 2

	Arousal				**Valence**			
	Delay	RMSE	CC	CCC	Delay	RMSE	CC	CCC
LBP	0	0.225	0.347	0.189	4.8	0.254	-0.037	-0.024
LPQ	0.4	0.226	0.341	0.273	2	**0.221**	0.039	0.026
LGBP	2.4	0.232	0.304	0.244	4.8	0.249	0.024	0.012
LBP-TOP	0.4	**0.224**	0.354	0.195	7.6	0.250	-0.054	-0.036
LPQ-TOP	1.2	0.228	**0.366**	**0.311**	1.2	0.228	**0.116**	**0.074**
LGBP-TOP	1.6	0.233	0.283	0.220	0.8	0.236	0.055	0.032

(c) Block size: 4 × 4

Table 5.5: Results on development partition after post-processing steps

	Arousal		**Valence**	
	Before	After	Before	After
RMSE	0.228	0.259	0.228	0.277
CC	0.366	0.431	0.116	0.133
CCC	0.311	0.399	0.074	0.122

quality of the data obtained from the Kinect. This is shown in Figure 5.7. The quality of such depth images often suffers from limited accuracy and stability due to invalid depth values (shown as black pixels in the Figure) and inconsistent depth values. The invalid depth values usually occurs on the boundary of objects, and smooth or shiny surfaces (e.g. the glass region as shown in the Figure). Furthermore, the depth value of a particular pixel can change from one frame to the next even when the scene is static.

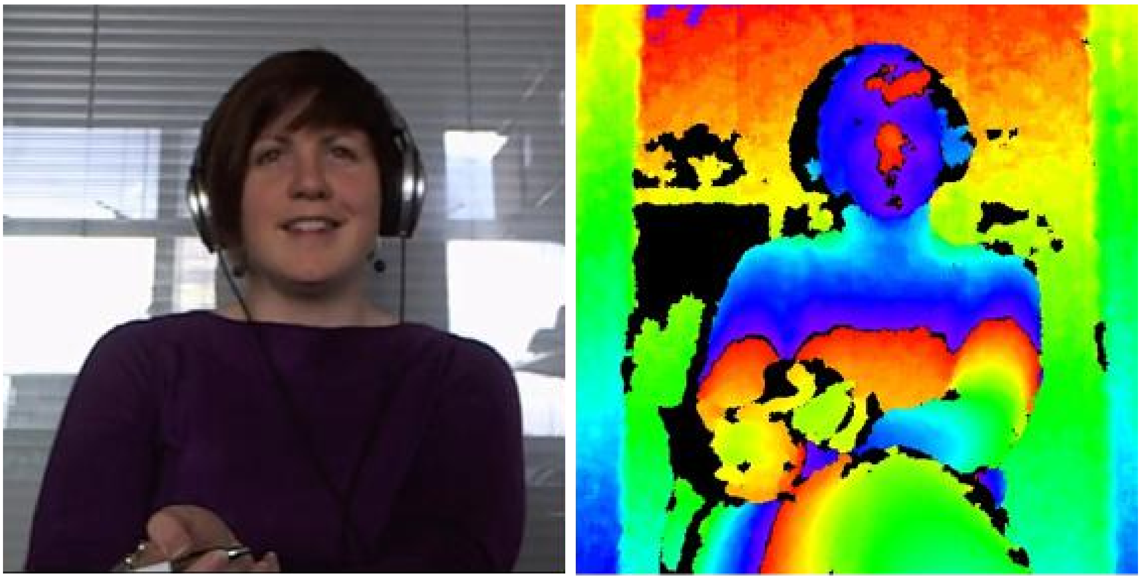


Figure 5.7: Data obtain from HD webcam and Kinect. Left: Colour image. Right: Depth Image

Figure 5.8 and Figure 5.9 shows the arousal and valence prediction results for the same recording using video and depth modalities respectively along with the ground truth. As can be seen for the arousal dimension, the predictions from video modality (red) are more correlated to the ground truth (blue) when compared to the predictions from the depth modality (yellow). However, there are instances where the depth modality gives a better prediction as marked by the dashed line. For the valence dimension, both video and depth modalities performed poorly at predicting the ground truth due to the unreliable annotation on the valence dimension as discussed in Section 3.5. This means that for similar input data, the annotation could be very different. As a result, the trained model might not be able to model the data correctly and thus give poor recognition results.

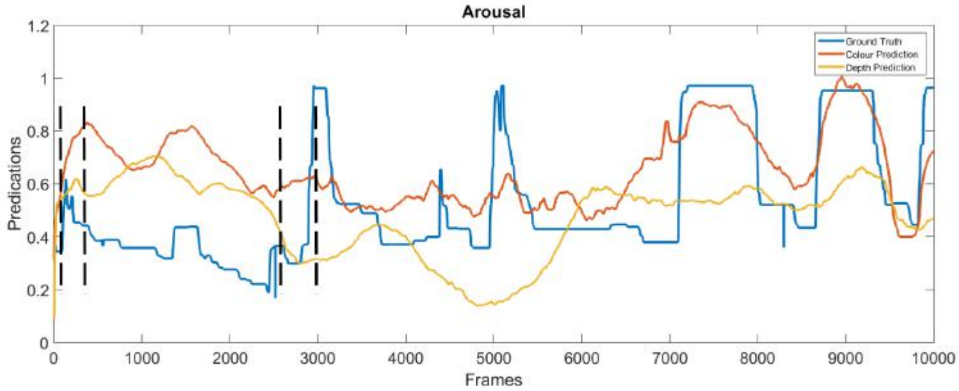


Figure 5.8: Comparison between colour and depth predictions on arousal dimension

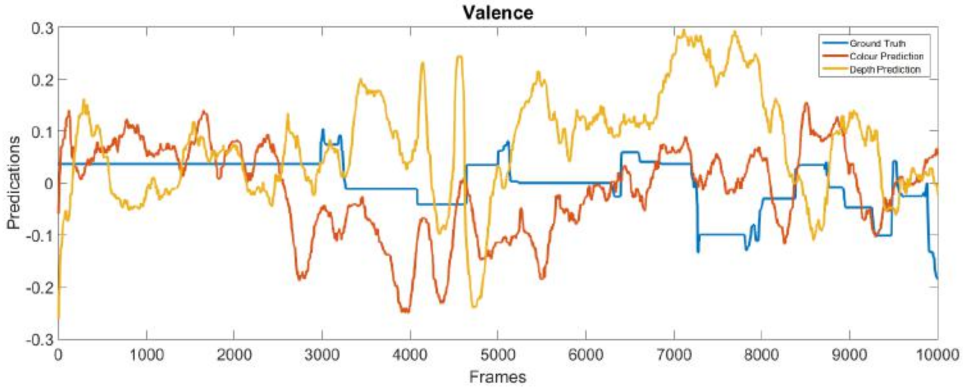
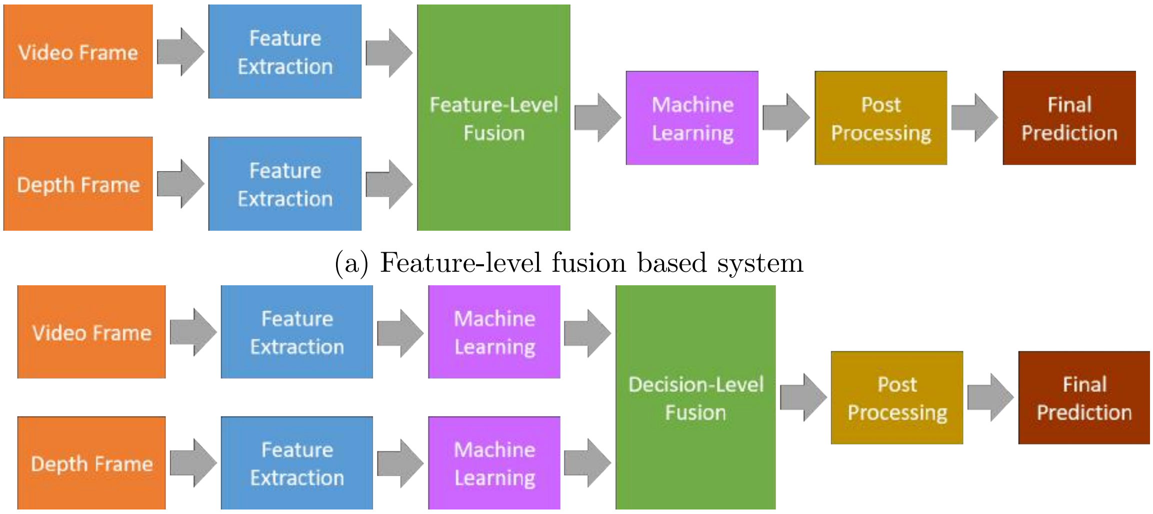


Figure 5.9: Comparison between colour and depth predictions on valence dimension

5.4 A Multi-Modal Framework

The previous experiments suggest that by combining video and depth modalities, the recognition results could be further improved when compared to using each modality alone. A multi-modal framework is proposed in this section, which use both modalities. During the training phase, the best histogram-based features are selected along with the annotation delay and the best window size for post-processing steps. When the system is in use, the face detection results from video and depth modality are compared. When the video modality is not visible (e.g. due to low light conditions), the depth modality is used solely for the prediction. When the face is detected for both modalities, the predictions can be made either through feature-level fusion or decision-level fusion.



(a) Feature-level fusion based system

(b) Decision-level fusion based system

Figure 5.10: Overview of feature-level and decision-level multi-modal system

5.4.1 Experiment 1: Feature-Level Fusion

The aim of this experiment is to evaluate the performance of feature-level fusion when applied to video and depth modalities.

5.4.1.1 Experimental Procedure

In this experiment, the best features and dimensions are selected for each modality based on the previous experimental results. For the video modality, the 1×1 LPQ-TOP feature is selected for the arousal dimension while the 4×4 LBP-TOP feature is selected to valence dimension. For the depth modality, the 4×4 LPQ-TOP feature is chosen for both arousal and valence dimensions. Before training, all features from different modalities are concatenated in a frame-by-frame manner. This results in 548 features for the arousal dimension and 287 features for the valence dimension.

Similarly to experiments described in Section 4.3, the annotation delay is shifted from 0 to 8 seconds with a time step of 0.4 seconds and the best time delay is selected using the development partition of the training dataset. A Support Vector Regression (SVR) is used to perform the regression task with the liblinear library. As the experiments carried out in Section 4.2, the L2-regularised L2-loss dual solver was chosen and a unit bias was added to the feature vector. During training, feature vectors containing all 0s were dropped. The complexity parameter c of the SVR was optimised in the range of $[10^{-5} - 10^{0}]$ on the development partition, while other parameters are kept as default. The post-processing steps proposed in Section 4.4 were then applied to the predictions after delay compensation.

5.4.1.2 Experimental Results and Analysis

Table 5.6 shows the comparison between the prediction results from single modalities and feature-level fusion. As can be seen, for the arousal dimension the feature-level fusion outperformed both video and depth modalities in terms of RMSE, CC and CCC scores. However, for the valence dimension the feature level fusion does not improve recognition results when compared to using the video modality alone. This could be caused by annotators giving different values for the same data on the valence dimension as discussed in Section 4.5.2.

5.4.2 Experiment 2: Decision-Level Fusion

The aim of this experiment is to evaluate the performance of decision-level fusion when applied to video and depth modalities. Two machine learning techniques were

Table 5.6: Comparison between single modality and feature-level fusion on development partition

	Arousal			**Valence**		
	Video	Depth	Fusion	Video	Depth	Fusion
RMSE	0.175	0.259	0.171	0.189	0.277	0.189
CC	0.777	0.431	0.791	0.258	0.133	0.246
CCC	0.725	0.399	0.739	0.178	0.122	0.169

investigated for the decision fusion including linear regression and LSTM.

5.4.2.1 Experimental Procedure

In this experiment, the predictions from Section 4.4 were used directly. In addition, the linear regression and LSTM were selected for decision-level fusion. For linear regression, the Weka 3.7 (Hall *et al.*, 2009) library with default parameters were used. For LSTM, the CURRENNT toolkit (Weninger *et al.*, 2015) was used. The LSTM network consists two hidden layers where each hidden layer consists of 100 and 80 LSTM units respectively. Due to the limitation of GPU memory, each recording is divided into multiple sequences where the maximum frames for each sequence is set to 1000 frames. To improve generalisation and prevent over-fitting, Gaussian noise with a standard deviation 0.1 is applied to all inputs. The weights are initialised randomly using uniform distortion between -0.1 and 0.1. The network is trained using mini batches of 10 sequences and for a maximum of 200 epochs. Training is stopped if no improvement in the performance (by RMSE) is observed on the development partition for more than 50 epochs. The prediction results from decision-level fusion were then post-processed with the same approach used in Section 4.4.

5.4.2.2 Experimental Results and Analysis

Table 5.7 shows the comparison between the recognition results using linear regression and LSTM respectively. Compared to the feature-level fusion approach, decision-level fusion has improved the recognition result on both arousal and va-

Table 5.7: Comparison between linear regression and LSTM recognition results on development partition

	Arousal		**Valence**	
	Linear Regression	LSTM	Linear Regression	LSTM
RMSE	0.170	0.169	0.181	0.176
CC	0.795	0.810	0.247	0.262
CCC	0.740	0.751	0.178	0.195

Table 5.8: Recognition results comparison between single modality, feature-level and decision-level fusion on test partition. FL : Feature Level Fusion, LR : Linear Regression

	Arousal					**Valence**				
	Video	Depth	FL	LR	LSTM	Video	Depth	FL	LR	LSTM
RMSE	0.187	0.267	0.185	0.184	0.183	0.210	0.281	0.211	0.199	0.195
CC	0.653	0.363	0.668	0.671	0.695	0.211	0.101	0.201	0.213	0.237
CCC	0.610	0.313	0.621	0.635	0.648	0.156	0.098	0.147	0.160	0.169

lence dimensions. In particular, the LSTM decision-level fusion method achieved the best results when compared to the linear regression approach.

Table 5.8 shows the recognition results using a single modality, feature-level fusion and decision-level fusion on the test partition of the DCU dataset. Unlike previous experiments where only the training partition was used to train the model, both training and development partitions were used to train the model when evaluated on the test partition. As can be seen, for both arousal and valence dimensions, the results obtained on the test set were quite similar to the development set, which means the proposed system could generalise well for unseen data. However, due to the unreliable annotation on the valence dimension as discussed in Section 4.5, the performance on the valence dimension was significantly lower when compared to the arousal dimension.

5.5 Conclusion

In this Chapter, experiments carried out to study the performance of the proposed continuous affect recognition system that uses the depth modality alone and both video and depth modalities are described. Before feature extraction, the face region needs to be located. In order for the depth modality to work independently from the video modality, a depth face detector was first proposed in Section 5.2. To train the depth face detector, 420 depth images were extracted from the dataset introduced in Chapter 3. The HOG feature and structural SVM was used to train the face detector and it achieved 97% accuracy when measured using a person independent 5-fold cross validation technique.

Section 5.3 evaluated the performance of the histogram-based depth features on continuous affect recognition. The results show that the histogram-based depth features can be used for continuous affect recognition. Among different features, the 4×4 LPQ-TOP feature achieved best recognition results on both arousal and valence dimensions. Similar to the features extracted from the video modality, the prediction was higher on the arousal dimension then the valence dimension which might be caused by the unreliable annotation on this dimension. However, compared to the video modality, the performance using depth was relatively low on both arousal and valence dimension which could be caused by the noise in the depth data obtained from the Kinect. By comparing the predictions from video and depth modality, it can be seen although the video modality result was better correlated with the ground truth, there were situations where the depth modality gave better predictions. This might suggest that by combining the video and depth modality, the performance of the system could be further improved.

In Section 5.4, a system has been developed and experiments have been carried out to study if the use of both video and depth modalities could lead to a improvement in the performance in affect recognition. In particular, feature-level fusion and decision-level fusion methods are investigated. The results show that feature-level fusion improved the recognition results on the arousal dimension, but not on the valence dimension, while decision-level fusion improved the results on both dimensions. However, due to the unreliable annotation on the valence dimension, the improvement is relatively small when compared to arousal dimension. The evaluation results on the test partition of DCU dataset have indicated that the proposed system could generalise well for unseen data.

Chapter 6

Conclusion

6.1 Overview

The research reported in this thesis examined the use of features from video and depth modalities for continuous affect recognition. In particular, histogram-based features are applied to colour and depth modalities. The best feature configurations are examined and evaluated on multiple datasets. A multi-modal affect recognition framework which utilises both colour and depth information is proposed. The following section reviews the key findings of each chapter along with suggestions for further works.

6.2 Thesis Summary

In **Chapter 1**, the motivation and research objectives associated with using the visual modality for continuous affect recognition of this thesis was introduced. The motivation for carrying out the research is based on the fact that current designs for Human Computer Interactions ignore the user's affective state, and automatic affect recognition is needed to enable a personal and more satisfying user experience in the future. A brief overview of research in the affect recognition area has identified that the current focus of automatic affect recognition research is to continuously recognise spontaneous affective state using multi-modal cues in valence-arousal dimensional space. It was found that although existing continuous affect recognition has proven to be highly successful using data captured in a controlled environment, little work have been done on using data from unconstrained environments. Following the brief overview of the research in affect recognition, four research objectives are proposed. These are:

1. Thoroughly investigate the performance of popular low-level appearance features for affect recognition.
2. Explore if individual facial parts such as mouth and eyes could be used to recognise affective state.
3. Investigate if data captured from a low-cost depth sensor could be used for continuous affect recognition.
4. Investigate if the use of both video and depth modalities could lead to a significant increase in performance in affect recognition.

In **Chapter 2**, the technical background necessary for understanding the research in this thesis was described. The review starts with an overview of the three most commonly used emotion models, namely the categorical model, the dimensional model and the appraisal model, along with their advantages and disadvantages. The dimensional model is chosen for the work carried out in this thesis. In particular, the arousal and valence dimensions are selected due to their widespread use in continuous affect recognition research. Then various techniques used to construct an affect dataset are discussed and the existing datasets in the literature are compared. It was found that although a number of datasets have been captured to fullfill the needs of training and testing automatic affect recognition systems, there still does not exist any dataset that includes recordings of spontaneous behaviours with 3D information that is also annotated continuously using the dimensional model. To address these issues a multi-modal multi-speaker 3D spontaneous affect dataset is needed. The details of constructing such a dataset are presented in Chapter 3. Various enabling aspects of affect recognition using the visual modality are then reviewed, including feature extraction, machine learning, multi-modal fusion and performance evaluation. The review shows that there are three areas that require further investigation. Firstly, among different low-level appearance features little research has been done to investigate their performance across different datasets. Secondly, features used in most of the existing affect recognition systems are holistic facial features, while no research has been carried out to study the use of facial parts features for continuous affect recognition. Thirdly, most of the existing research focuses on using 3D visual features to predict discrete emotions and no research has been carried out on using them for continuous affect recognition. In the last section of this chapter, a number of baseline systems are introduced to benchmark the approaches proposed in this thesis.

In **Chapter 3**, a multi-modal data capture platform that can be used to capture video, audio and depth simultaneously was presented. The developed platform was used to capture one of the first multi-modal multi-speaker debate affect dataset. In total 16 participants and over 5 hours of data were recorded. GTrace was then employed to annotate the data continuously along 5 dimensions including arousal, valence, agreement, content and interest. To evaluate the reliability of the annotations, statistical analysis was performed using three metrics including the percentage of positive frames, the mean Correlation Coefficient (CC) and Cronbach's α. The analysis of the annotations indicates good inter-agreement on the arousal and interest dimensions and acceptable inter-agreement on the valence, agreement and content dimensions, making it a suitable dataset to evaluate the performance of the systems proposed in the following chapters.

In **Chapter 4**, a number of experiments were carried out to explore the use of histogram-based appearance features extracted from the video modality for continuous affect recognition. The first set of experiments thoroughly investigated the performance of histogram-based features and the best configurations when they are used to predict affective state. The experimental results show that the LBP-TOP and LPQ-TOP features in general perform best for continuous affect recognition. However, the best block size configurations tend to vary for different datasets and dimensions when measured in terms of CC score. The annotation delay analysis shows that the annotation delay also tends to vary for different datasets and dimensions. In general, the delay is longer for the valence dimension than the arousal dimension. The second set of experiments investigated if individual facial parts could be used instead of the whole face region to predict affective state. Experimental results indicate that by using Convolutional Neural Networks (CNN) it is not only possible to predict affective state using facial parts, but also that it achieved better results when compared to the use of the entire face image directly as input. In general, the eye region is good at predicting arousal while the mouth region performs best for valence prediction.

In **Chapter 5**, the challenges encountered when creating a multi-modal affect recognition system were explored and addressed . Firstly, a face detector based on a depth image is proposed in order for the video and depth modalities to work independently. The results show that the use of depth image achieved comparable results as using the colour image. Secondly, experiments are conducted to study the performance of the histogram-based features when applied to the depth modality. The results show that the histogram-based depth features can be used for continuous affect recognition. Among different features, the 4×4 LPQ-TOP feature achieved best recognition results on both arousal and valence dimensions when evaluated on the DCU dataset. Thirdly, the use of feature-level and decision-level fusion methods to fuse different modalities are examined and a multi-modal affect recognition framework is proposed. The results show that feature-level fusion improved the recognition results on the arousal dimension but not on the valence dimension, while decision-level fusion improved the results on both dimensions.

6.3 Analysis and Discussion of Research Objectives

In this thesis, a number of research objectives are explored to investigate how visual modality can be used for human affective state recognition. In this section, the research objectives are examined with respect to the experimental results obtained.

- **Research Objective 1**

  Thoroughly investigate the performance of popular low level appearance features for affect recognition.

  The first four experiments conducted in Chapter 4 explored the use of the six most commonly used histogram-based features including LBP, LPQ, LGBP, LBP-TOP, LPQ-TOP and LGBP-TOP for continuous affect recognition. Results show that in general the dynamic (TOP) features achieved better CC scores on both arousal and valence dimensions compared to the static features. In addition, the LBP-TOP and LPQ-TOP features perform best for continuous affect recognition. However, the best block size configurations tend to vary for different datasets and dimensions when measured in terms of CC score. The experimental results also show that by introducing delay compensation and post-processing steps the recognition results could be further improved in terms of both CC and CCC scores. The annotation delay also tends to vary for different datasets and dimensions. In general, the delay is longer for the valence dimension than the arousal dimension. By combining the best histogram features with Support Vector Regression (SVR) and post-processing techniques, the proposed system has achieved better results on the valence dimension and comparable result on the arousal dimension in terms of both CC score when compared to the AVEC 2012 and AVEC 2015 winning systems.

- **Research Objective 2**

  Explore if individual facial parts such as mouth and eyes could be used to recognise affective state.

  This research objective is addressed in the final section of Chapter 4. The experiments on individual facial parts have shown that by using Convolutional Neural Networks (CNN) it is not only possible to predict affective state using facial parts, but also that it achieved better results when compared to the use of the entire face image directly as input. The results show that the eye region is good at predicting arousal while the mouth region performs best for valence prediction. The annotation delay analysis suggests that, when annotating the data, the reaction time for different dimensions is different. The mouth region requires a longer reaction time when judging arousal than valence, and the nasolabial folds require longer reaction time for valence than arousal. Compared to histogram-based features, the use of facial parts achieved comparable results on the valence dimension, but lower results on the arousal dimension, which might due to the fact that the amount of training data is relatively small for the CNN to learn representative features.

- **Research Objective 3**

  Investigate if data captured from low cost depth sensor could be used for continuous affect recognition.

  This research objective is explored in Chapter 5, where the performance of the histogram-based depth features for continuous affect recognition are evaluated. The results show that the histogram-based depth features can be used for continuous affect recognition. Among different features, the 4 × 4 LPQ-TOP feature achieved best recognition results on both arousal and valence dimensions. By comparing the predictions from video and depth modality, it can be seen that although the video modality result was better correlated with the ground truth, there were situations where depth modality gave better predictions. This suggests that by combining the video and depth modality, the performance of the system could be further improved.

- **Research Objective 4**

  Investigate if the use of both video and depth modalities could lead to a significant increase in performance in affect recognition.

  The last section in Chapter 5 investigates this research objective. In particular, feature-level fusion and decision-level fusion methods are explored. The results show that feature-level fusion improved the recognition results on arousal dimension but not on valence dimension while decision-level fusion improved the results on both dimensions. However, the improvement for both methods are relatively small which could be caused by the unreliable annotations of the ground truth. Finally, a multi-modal framework which can be used in real world settings is proposed based on the experimental results.

6.4 Further Work

The experiments conducted in this work lead to several conclusions which pave the way to further research. In this section, these potential research areas which could add to the literature in automatic affect recognition are analyzed.

- **Cross-dataset learning:** As discussed before, deep neural networks have shown superior performance on various machine learning tasks. However, for deep neural networks to learn representative features, a large amount of training data is required. With the availability of various affect datasets as discussed in Section 2.3.4, it would be interesting to explore the combinations of multiple datasets which might improve recognition results.

- **3D face modelling:** The work described in this thesis has been mainly focused on using low level features derived directly from the video and depth images. Future research can be carried out to explore the use of features extracted from 3D facial models. The 3D facial models can be reconstructed from both video and depth data which are robust to large changes including out-of-plane head rotations, fast head motions and partial facial occlusions.

- **Multi-modal feature fusion:** As stated earlier, the current trend in automatic affect recognition focuses on using multiple modalities. The works described in this thesis has investigated the use of video and depth modalities. However, the audio modality could be utilised to enhance the performance of the proposed system. This thesis only focused on using histogram-based features whereas other geometric and appearance-based features have also shown good performance on continuous affect recognition. Hence, it will be interesting to explore the fusion of multiple feature representations to improve the efficacy of recognition.

- **Exploring alternative machine learning methods:** Further research can be carried out to investigate other machine learning techniques viz., ensemble learning and multi-task learning. Ensemble learning is a machine learning method where multiple learners are used to solve the same problem. It usually provides a much stronger generalisation ability when compared to conventional approaches. Multi-task learning is a another learning paradigm that aims to utilise useful information to learn various related tasks simultaneously to improve the generalisation performance, i.e. instead of learning arousal and valence dimension separately, one can use the multi-task learning method to learn the arousal and valence at the same time.

Appendix A

Baseline System Implementation

A.1 Introduction

As briefly discussed in Chapter 2, various multi-modal affect recognition systems have been proposed in the literature. For the research reported in this thesis it is necessary to choose a baseline system so that the performance of the proposed framework on various databases can be benchmarked. In this Appendix, different existing frameworks are first compared and the most appropriate one is chosen as the baseline system. Following this, the implementation details of the baseline system are discussed and its performance is analysed.

A.2 Comparison of Existing Systems

To select the most appropriate baseline system, a number of systems from the Audio-Visual Emotion Challenge (AVEC) are chosen. AVEC is an annual competition event aimed at automatic affect analysis. The challenge provides a common benchmark dataset for multimodal affect recognition. In particular, the systems from AVEC 2012 (Schuller *et al.*, 2012) are compared since this was the most recent system at the time of implementation.

Table I: AVEC 2012 Audio Low-Level Descriptors (LLD) Schuller *et al.* (2012)

Energy & Spectral (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250-650 Hz, 1kHz-4kHz, 25%, 50%, 75% , and 90% spectral roll-off points spectral flux, entropy, variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity, MFCC 1-10
Voicing Related (6)
F_0 (sub-harmonic summation, followed by Viterbi smoothing) probability of voicing, jitter, shimmer (local), jitter (delta: ”jitter of jitter”), logarithmic Harmonics-to-Noise Ratio (logHNR)

The system proposed by Schuller *et al.* (2012) utilises Local Binary Pattern (LBP) features as the video features and statistical features of the low level descriptors for audio features (as shown in Table I and Table II). The features are then learned using Support Vector Machine regression (SVR) with Histogram Intersection Kernels and a Sequential Minimal Optimization (SMO) technique. This system was used as the baseline system for the AVEC 2012 challenge.

The system proposed by Nicolle *et al.* (2012) uses the log-magnitude Fourier spectra to extract dynamic information from the signal that describes the shape deformation, local and global face appearance. The same set of features as in Schuller *et al.* (2012) are used for audio features. A correlation-based feature selection process is then applied to select a relevant set of features, followed by a weighted K-Means and Nadaraya-Watson kernel regression. As the final step the predictions from each feature set are fused using a local linear regression to produce the final prediction.

Another work proposed by Savran *et al.* (2012a) uses Bayesian filtering with particle filtering to combine the features extracted from the video, audio and lexical modalities. The video features are extracted using Local Binary Patterns (LBP) based on temporal statistics, while the audio features include a subset of features

Table II: Set of all 42 functionals. [1]Not applied to delta coefficient contours. [2]For delta coefficients the mean of only positive values is applied, other wise the arithmetic mean is applied. [3]Not applied to voicing related LLD Schuller *et al.* (2012)

Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1%, 99% percentile, percentile range 1%99%, percentage of frames contour is above: minimum + 25%, 50%, and 90% of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[1,3], standard deviation of segment length[1,3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1,3] (6)
LP gain, LPC 1-5

used in Schuller *et al.* (2012) plus class-level spectral features based on three distinct phoneme classes. The lexical features are calculated as the pointwise mutual information (PMI) between a word and a given affect dimension. A Support Vector Machine for Regression (SVR) is then used for each modality and the final results are fused using a Bayesian framework via particle filtering.

The work reported by Baltrusaitis *et al.* (2013) proposed a framework to utilise the combination of Continuous Conditional Random Fields (CCRF) and SVR for modeling continuous affective state in dimensional space. For video features, the system extracts the geometric features described by the expression parameter, along with appearance features described by Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) and motion features described by head movements. The prosodic features used in Ozkan *et al.* (2012) are adopted as audio features. SVR is then used to predict the affective state for each of the four feature sets (geometric, appearance, motion and audio) and the final results are fused using the CCRF.

The recent work by Wei *et al.* (2014) developed a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) and multiple kernel learning (MKL) based multi-modal affect prediction framework (LSTM-MKL). Their motivation was to leverage the advantages of LSTM-RNN for modeling long range dependencies between observations and MKL for modelling non-linear correlations between input and output. The system uses visual features proposed in Savran *et al.* (2012a) and audio features detailed in Schuller *et al.* (2012).

The prediction results measured in terms of Pearson Cross Correlation (See Section 2.4.4 for more details) of the above systems are shown in Table III. The text in bold indicates the highest score for a particular dimension. Compared to the benchmark system proposed by Schuller *et al.* (2012), all four systems showed a significant increase across all four dimensions. This could be a result of introducing information on previous affective states and across dimensions. Among the four systems, the system developed by Nicolle *et al.* (2012) achieved the best arousal, expectancy

Table III: Pearson's correlation score for different systems tested on AVEC 2012 development database

System	**Aro**	**Exp**	**Pow**	**Val**	**Mean**
Schuller 2012 Schuller *et al.* (2012)	0.181	0.148	0.084	0.215	0.157
Nicolle 2012 Nicolle *et al.* (2012)	**0.644**	**0.341**	0.511	0.350	**0.461**
Savran 2012 Savran *et al.* (2012a)	0.383	0.266	**0.556**	**0.473**	0.384
Baltrusaitis 2013 Baltrusaitis *et al.* (2013)	0.333	0.218	0.309	0.343	0.301
Wei 2014 Wei *et al.* (2014)	0.453	0.298	0.339	0.327	0.354

and average prediction result. It also achieved the second best result on valence and very close to best result on power. In addition, part of the feature extraction code used in the system is publicly available. It is proposed here that this system is both the best performing and most reproducible baseline system. For the above reasons, the system proposed by Nicolle *et al.* (2012) was selected as the baseline system. Although this thesis focused on using visual features for affect recognition, in order to compare the results from the implementation with the original paper the audio features are also used for consistency

A.3 Baseline System Implementation

As shown in Figure I, the baseline system consists of three parts: feature extraction, affect prediction and result fusion. This section first gives a general description of each of these components and implementation details are then discussed.

A.3.1 Feature Extraction

In this section the feature extraction component of Figure I is explained. The baseline system uses four different sets of features. The first three sets of features are based on visual cues while the fourth one is based on audio. This is shown in Figure II. The visual features include shape parameters, global face appearance and local face appearance.

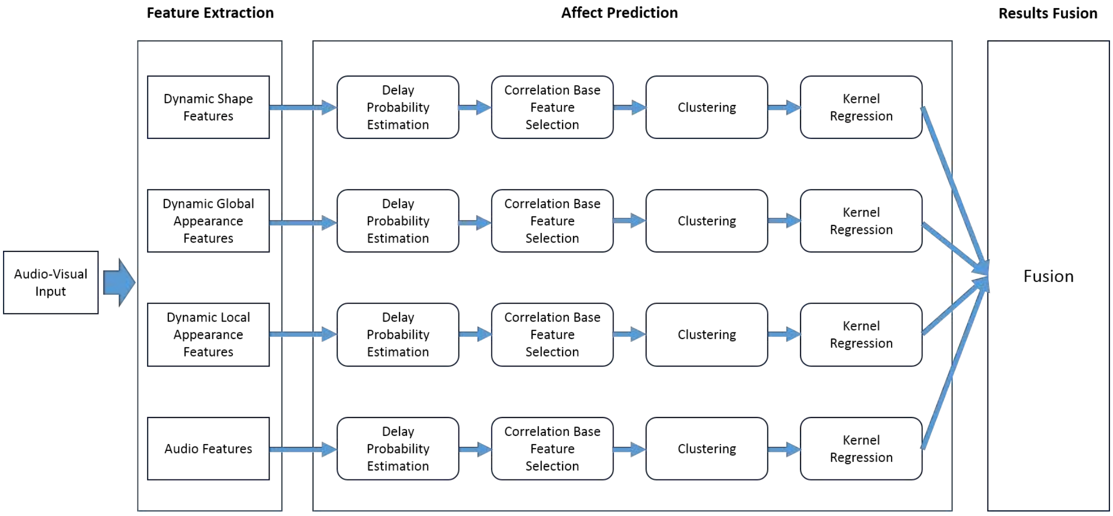


Figure I: Overview of the baseline system

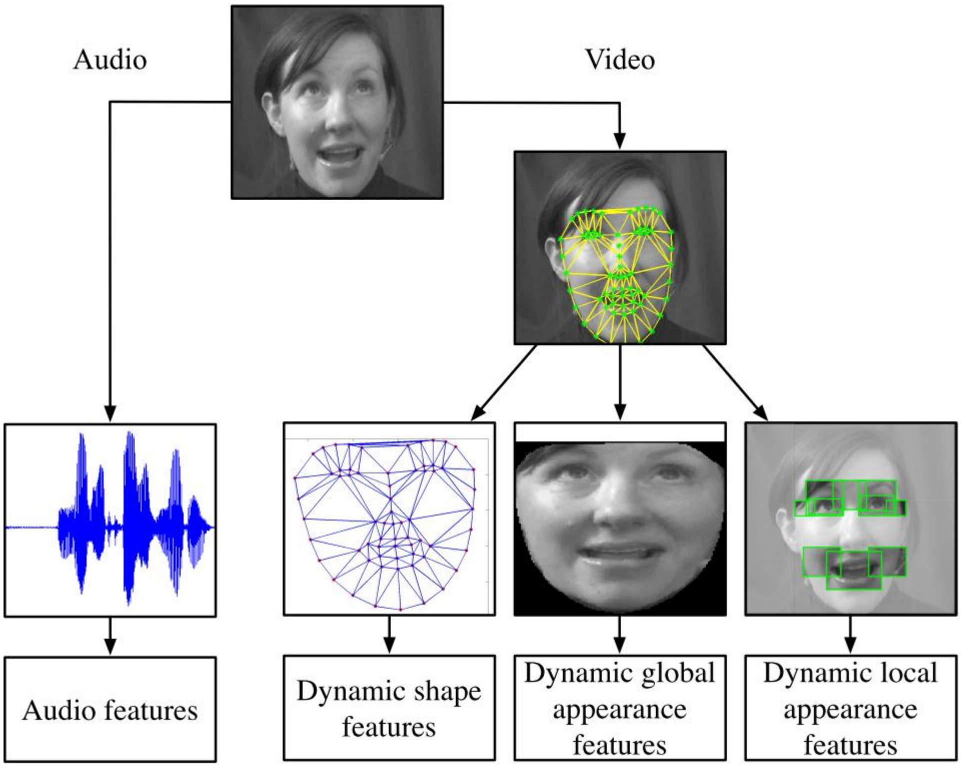

Audio
Video
Audio features
Dynamic shape features
Dynamic global appearance features
Dynamic local appearance features

Figure II: Baseline system feature set (Nicolle *et al.*, 2012)

A.3.1.1 Static and Dynamic Features

The three sets of visual features are first extracted for each frame. Then a dynamic descriptor for each of the visual feature sets is extracted using different window sizes as discussed in Section A.3.1.5.

A.3.1.2 Shape Features

As discussed before, the face shape can be represented by a set of suitable landmarks Cootes (See Figure 2.16). Landmarks are defined as important geometric features for faces. They have been well studied and are used widely in biology and statistical shape analysis (Dunn, 1993; Dryden and Mardia, 1998). As defined by Dryden and Mardia (1998), "A landmark is a point of correspondence on each object that matches between and within populations".

There are three types of landmarks that can be used to model the face. The *anatomical landmarks* are defined by experts and have biological meaning, e.g. corners of eyes and mouth. The *mathematical landmarks* are defined according to mathematical or geometric properties, e.g. centre point on outer edge of lower or upper lip. Finally *pseudo-landmarks* are usually defined as equally spaced points between anatomical or mathematical landmarks or outline of the face, e.g. contours of the cheek (Shi *et al.*, 2006).

A shape model is used to capture the structure and variations among the annotated landmarks by linearly combining a base shape $\overline{X}$ with n shape vectors X_i as shown in equation A.1. The base shape $\overline{X}$ and shape vector X_i are normally computed from hand-labelled training images. The training images are first aligned using Procrustes analysis (Cootes *et al.*) to remove the global rigid motion. Mathematically, this can be interpreted as simultaneously finding a canonical shape and transforming each training shape into alignment with the canonical shape using a similarity transformation (Baggio, 2012). Then Principle Component Analysis

(PCA) is applied to the aligned shapes to obtain the base shape (mean) $\overline{X}$ and the shape vectors (eigenvectors) X_i. By varying the shape parameters different global and local variations can be modeled as shown in Figure III.

$$X = \overline{X} + \sum_{i=1}^{n} p_i X_i \tag{A.1}$$

where p_i are the shape parameters that represent the weight of the shape vectors.

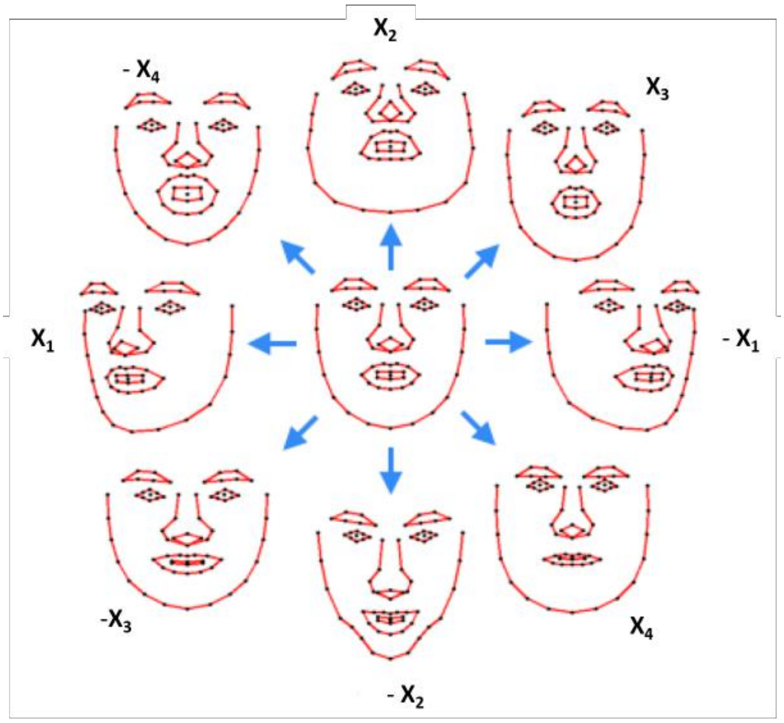


Figure III: Effects of varying each of the first four shape parameters X_i (Baggio, 2012)

After adding the similarity transformation parameters $T_{s,\ \theta,\ t}$ to equation A.1 the equation becomes:

$$X = T_{s,\theta,t}(\overline{X} + \sum_{i=1}^{n} p_i X_i) \tag{A.2a}$$

$$X = s\theta(\overline{X} + \sum_{i=1}^{n} p_i X_i) + t \tag{A.2b}$$

where the similarity transformation includes the translation t, scaling s and a rotation θ.

The baseline system uses the 3D face tracker proposed by Saragih *et al.* (2011) to detect and track the landmarks in the images. The paper uses shape features corresponding to the "external parameters" and "characterise deformations related to facial expression", however, no details are provided on what parameters from the face tracker are used and how frequently the features are extracted.

In the implementation described in this thesis, the transformation and shape parameters ($t,\ s,\ \theta,\ p_i$) from the face tracker are used. These parameters give information on the rigid and non-rigid transformation of the face. The shape features are extracted for each frame to form the shape feature vectors.

A.3.1.3 Global Appearance Features

Similar to the shape model, the appearance of the face A could also be modeled by a base appearance $\overline{A}$ plus a linear combination of m appearance images A_i as shown in equation A.3. In order to compute $\overline{A}$ and A_i, the hand-labeled training images are first warped onto the base shape s_0 using triangulation and a piecewise affine warp. Then PCA is applied to the *shape normalised* training images. $\overline{A}$ is set to be the mean image while A_i are set to be the m eigenimages corresponding to the m largest eigenvalues obtained from PCA. Figure IV shows how the face image could be represented using base appearance and appearance parameters.

$$A(x) = \overline{A}(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \qquad \forall x \in s_0 \tag{A.3}$$

where λ_i are the appearance parameters.

To build the global appearance model, the baseline system uses a number of images from the training set of the AVEC 2012 database. The landmarks of the face images are first detected using the same face tracker as mentioned in Section A.3.1.2. Then the important appearance modes are selected using PCA as described above.

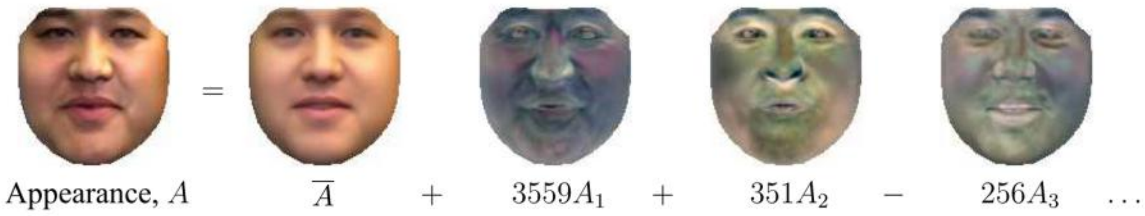


Figure IV: Appearance Model (Matthews and Baker, 2004)

By projecting the face image into the PCA space, the appearance parameters λ_i are obtained. However, in the paper there are no details on how many images are used to build the appearance model, how these images are selected, the image size and the amount of appearance variation in PCA space the model needs to cover.

In our implementation, in order to build the appearance model the images are extracted from video recordings of the training partition every 5 seconds since adjacent frames can consist of very similar facial expressions. This results in a total of 2000 face images that cover a range of head poses and affective states. The images are then scaled to 195×145 since this provides a balance between clear visibility of the face and the speed of the model building process. The appearance models are chosen to cover 98% of the appearance variations from the PCA result. By projecting the warped image to the appearance, a set of appearance parameters can be obtained. The appearance parameters are then calculated for each frame of the video recording to form the global appearance feature vectors.

A.3.1.4 Local Appearance Features

The local face appearance is defined as a set of local patches that involve facial actions such as smile and gaze direction. The baseline system extracted 6 types of local patches as shown in Figure II. These include areas around the mouth, eye, eyebrows, periorbital lines, glabellar lines, nasolabial folds and smile lines. However, similar to the global appearance features there are no details on how the local appearance model is built.

Table IV: Scale size for different local patch. PL: Periorbital Lines, GL: Glabellar Lines, NL: Nasolabial Folds, SL: Smile Lines

Patch Type	Mouth	Eyes	Eyebrows	PL	GL	NF and SL
Patch Size (w*h)	60×40	40×20	65×25	30×40	40×30	45×50

In the implementation here, the same set of training images used to build the global appearance model are used to build the local appearance model. The patches for each facial part are defined with respect to the landmarks. Each patch is then scaled to the same size based on its type as shown in Table IV. Again the sizes are chosen to provide a balance between visibility and the speed of the modelling process. PCA is then applied to the patch to compute important local appearance modes for each patch type. The appearance modes were selected to cover 98% of the local appearance variation. The local appearance parameters are calculated for each frame by projecting the local patch to the local appearance modes similar to the global appearance parameters. The local appearance parameters for each patch are then concatenated together to form the final feature vectors for that frame.

A.3.1.5 Dynamic Features

In order to encode the dynamic information of the visual features, the baseline system uses the log-magnitude Fourier spectra. The Fourier coefficients are calculated every 1 to 4 seconds with a step of 1 second. The frequency is also binned every 5Hz to reduce the noise and dimensions of the dynamic features. Other statistical features including mean, the standard deviation, the global energy and the first and second-order spectral moments are also calculated for each window and concatenated to form a feature vector for that instance. The corresponding ground truth label value is calculated as the average label value of the 4 seconds time window. This results in approximately 8500 training samples for each of the first three visual feature sets.

A.3.1.6 Audio Features

The baseline system utilises the audio features from the AVEC 2012 Challenge directly as shown in Table I. The audio features are extracted during speech with a 2 second sliding window, with a 0.5 second interval Schuller *et al.* (2012). In our implementation, the audio features are also used directly.

A.3.1.7 Feature Normalisation

As the paper suggested, the set of four features is normalised by subject in order to reduce the inter-subject variability and give the same value range (-1,1) for different features.

A.3.2 Affect Prediction

The prediction system shown in Figure I is trained for each of the four feature sets. First, the features are selected based on the correlation score between each feature and each time-delayed label dimension. Then the selected features are clustered into groups based on the feature weight to produce the representative samples. Kernel regression is then used to predict the label values based on the representative samples. The following section gives more information and implementation details for each component.

A.3.2.1 Annotation Delay Probability Estimation

Because the affect data is annotated continuously using the Gtrace tool (See Section 2.3.3 for more details), this introduces a delay between the frame being annotated and its corresponding label due to the reaction time of human annotator. In order to address this issue the baseline system proposed to estimate the delay probability of the label. The probability of different delay time ,$P(\tau)$, is calculated as shown in

Equation A.4

$$P(\tau) = \frac{1}{A}\sum_{i=1}^{n} \rho(f_i(t), y(t-\tau)) \tag{A.4}$$

where $y(t-\tau)$ is the the ground truth label time series $y(t)$ shifted forward by τ seconds, $f_i(t), i \in [1, n]$ is a set of n features, and ρ is the Pearson correlation coefficient between two time series:

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma Y} \tag{A.5}$$

where σ is defined as the standard deviation and μ refers to the mean. A is the normalisation term defined as:

$$A = \int_{-\infty}^{\infty} \sum_{i=1}^{n} \rho(f_i(t), y(t-\tau)) d\tau \tag{A.6}$$

$$A = \sum_{\tau=0}^{20} \sum_{i=1}^{n} \rho(f_i(t), y(t-\tau)) \tag{A.7}$$

In our implementation, the label of each dimension is delayed between 0 to 20 seconds as the baseline paper suggested. Furthermore, for this implementation we set the delay step to 1 second. The integration in Equation A.6 can then approximate by the summation in Equation A.7. Because the data is recorded in different video sequences, the delay probability is calculated as the mean of the delay probability for each video. During the implementation it was found that using all features that have very low correlation between the labels could corrupt the delay probability estimation. Given the lack of details presented in the paper describing how to address this problem, the top 200 most-correlated features are selected to estimate the delay probability as this produces the most similar results to those reported in the paper. Figure V shows a comparison of the delay probability distributions between the paper and our implementation for the shape parameter features of the training database in AVEC 2012. The implementation shows a similar delay probability for arousal, expectancy and valence. This agrees with the paper for an average delay between 3 and 4 seconds for valence and arousal, and between 5-6 seconds delay for expectancy and power. However the shape of the delay probability for the power dimension is quite different compared to the original paper. This might be due to the fact that the method used to select features to calculate the delay probability is different.

A.3.2.2 Correlation-Based Feature Selection

After estimating the delay probability for each dimension, a feature weight is calculated using the following formulas:

$$w(f_i(t), y(t)) = \int_{-\infty}^{\infty} \rho(f_i(t), y(t-\tau))P(\tau)d\tau \tag{A.8}$$

$$w(f_i(t), y(t)) = \sum_{\tau=0}^{20} \rho(f_i(t), y(t-\tau))P(\tau) \tag{A.9}$$

where P(τ) is the delay probability computed in the last step. After setting the delay time τ to vary between 0 to 20 seconds with a step of 1 second, the integration in Equation A.8 can then approximate by Equation A.9. Similar to the delay probability estimate, the feature weight is also calculated as the mean for different video sequences. This gives a measure of how the i_{th} feature is correlated with the labels and how consistent this is across different videos. The paper suggested to select features based on the feature weights that maximise the total feature score. However, no details are presented on how this is to be achieved. For our implementation, a threshold of 0.2 is used to filter out the un-correlated features.

A.3.2.3 K-Means Clustering

The training samples are then clustered into 60 groups to produce the representative samples as the paper suggested. Unlike traditional K-Means, the baseline system

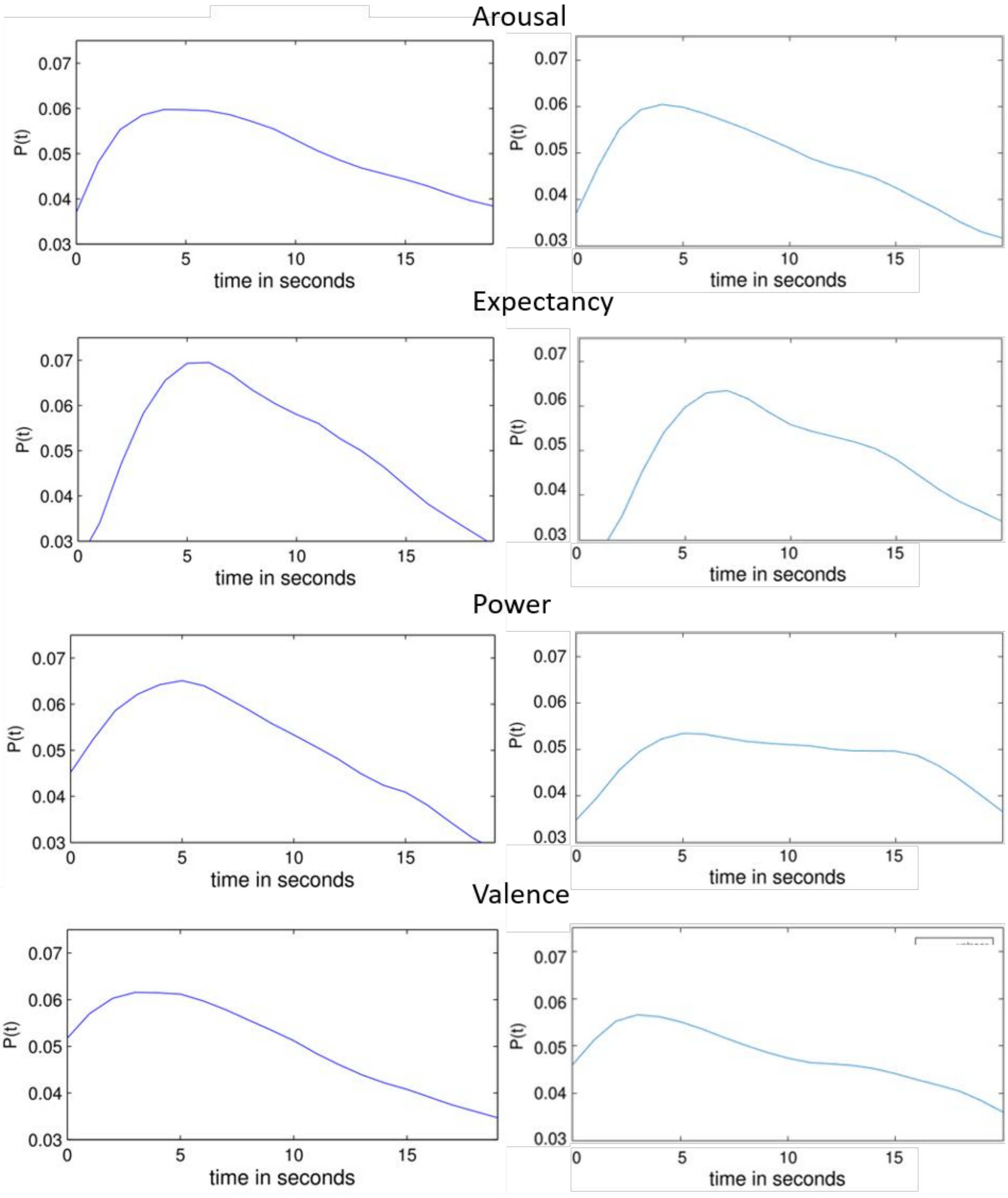


Figure V: Comparison of delay probability between baseline paper (left) and our implementation (right)

proposed a diagonally-weighted distance by using the feature weights as follows:

$$d_w(X, Y) = \sqrt{(X - Y)^T W (X - Y)} \tag{A.10}$$

where X and Y are two training samples, W is a diagonally weighted matrix consisting of feature weights defined as:

$$W_{ij} = w(f_i(t), y(t))\delta_{ij} \tag{A.11}$$

where $w(f_i(t), y(t))$ is the correlation score between each selected feature and the label, and δ_{ij} is the identity matrix.

Instead of selecting the initial centroid randomly, the training samples are first sorted according to each label's dimensions and divided into k groups of equal size. The initial centroids are then calculated as the mean of each feature in each group. After the clustering step, a set of k representative samples are produced. The label for each sample is computed as the mean label of all the training samples in that cluster. In our implementation the cluster size is selected based on the experiments discussed in Section A.4.1

A.3.2.4 Kernel Regression

The baseline system uses Nadaraya-Watson kernel regression (Nadaraya, 1964) to predict the labels. Kernel regression is a non-parametric model. It is used to estimate a regression function that best matches the training data. Compared to linear regression or polynomial regression, kernel regression does not assume any underlying distribution for the estimation of the regression function. Kernel regression positions a set of identical weight functions called the *Kernel Function* at each observed data point. The kernel function will assign a weight to each location based on the distance from each of the observed data points. Let x_i be the feature vectors of the k representative samples computed from the clustering step and y_i be the corresponding labels. The label value for a sample s with feature vector x_s is described by the following formula:

$$\hat{y}(s) = \frac{\sum_{i=1}^{k} K(x_s, x_i) y_i}{\sum_{i=1}^{k} K(x_s, x_i)} \tag{A.12}$$

where *K(x_c, x_i)* is the radial basis function (RBF) combined with the learned weighted-distance d_w defined as:

$$K(X_c, X_i) = e^{-\frac{d_w(x_s, x_i)^2}{2*\sigma^2}} \tag{A.13}$$

where σ is the kernel width to indicate how spread out each kernel is at each observed data point.

After the prediction step, the baseline system also performed temporal smoothing to reduce the noise of the prediction. In our implementation a moving average with window size 5 is used to post-process the prediction results.

A.3.3 Fusion

The prediction system is trained for each of the four feature sets as described in Section A.3.1 for each dimension resulting in 16 signals. The baseline system uses local linear regressions to fuse the signals and produce the final prediction. First the linear regression parameters for the i^{th} training video sequence and j^{th} dimension can be obtained by minimising the difference between the predicted value H_i and the ground truth Y_i^j. This can be solved in closed form by taking the derivative of the right hand side and setting to 0. This gives:

$$\theta_i^j = (H_i^T H_i)^{-1} H_i^T Y_i^j \tag{A.14}$$

The parameters α_j of the final linear regression model are computed as the means of θ_i^j weighted by the Pearson's correlation between the predicated signal and the ground truth of each video sequence as follows:

$$\alpha_j = \frac{\sum_{i=1}^{n} r(H_i\theta_i^j, Y_i^j)\theta_i^j}{\sum_{i=1}^{n} r(H_i\theta_i^j, Y_i^j)} \tag{A.15}$$

The final prediction for the four dimensions can then be computed as;

$$\hat{y_j} = H_t\alpha_j \tag{A.16}$$

where H_t contains the prediction of 16 signals of the test data.

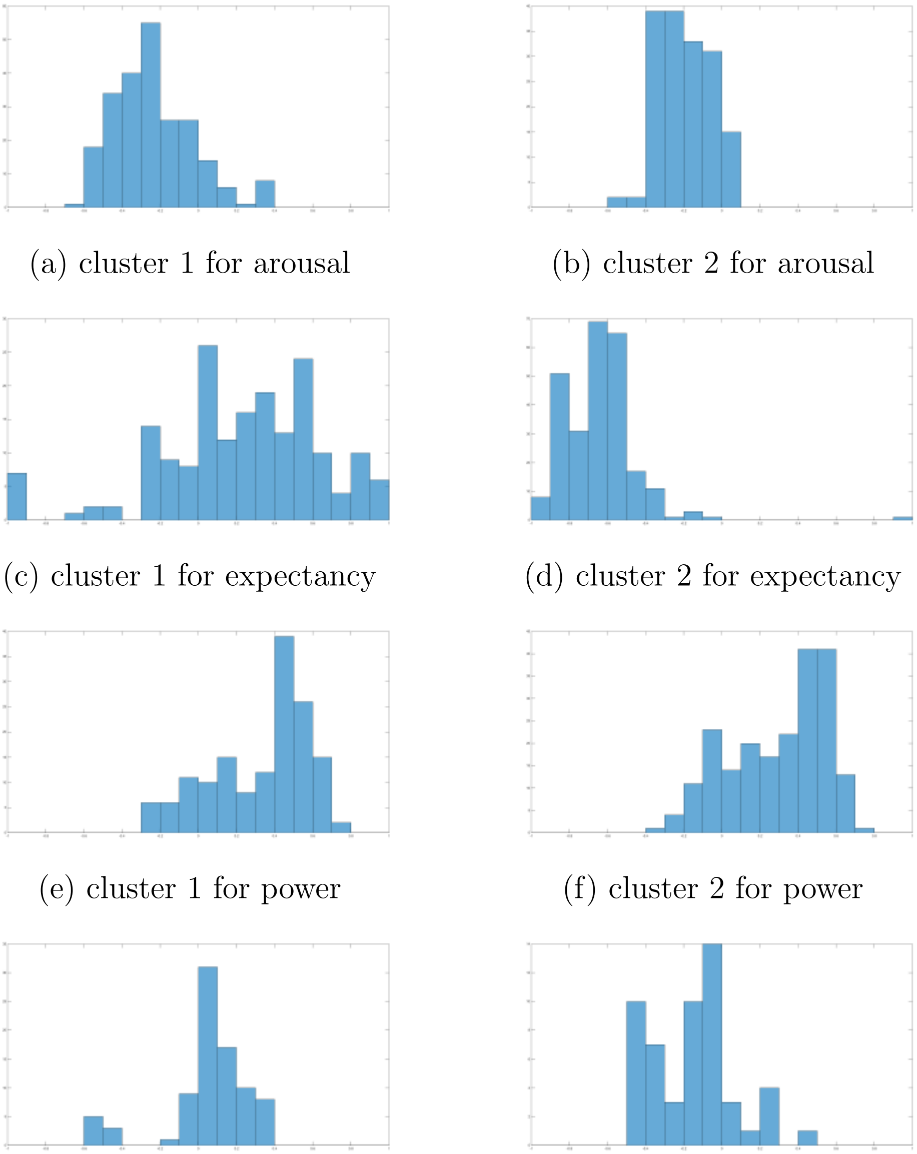
A.4 Experiments and Results

In this section, experiments carried out to optimise the parameters of the baseline system are described. This includes investigating the cluster size for the clustering step, the kernel width σ for kernel regression and the hyperparameters for the local linear regression. Finally, the results from the implemented baseline system are compared against the original paper.

A.4.1 Optimising Cluster Size

The goal of the clustering step described in Section A.3.2.3 is to group the training samples that have similar feature values together based on the assumption that these training samples also have similar affect values in each dimension. This assumes that the selected features are sufficiently discriminative to separate samples in different affective ranges. In the worst case, only a fraction of the samples in each cluster would follow a *priori* probability of the training data.

During the implementation it was found that the label values in each cluster are



(a) cluster 1 for arousal

(b) cluster 2 for arousal

(c) cluster 1 for expectancy

(d) cluster 2 for expectancy

(e) cluster 1 for power

(f) cluster 2 for power

(g) cluster 1 for valence

(h) cluster 2 for valence

Figure VI: The affect value distribution for the first two clusters (60 clusters in total) of each dimension for dynamic shape features

Table V: Average label variance for each affect dimension with different cluster sizes using dynamic shape features

No. of Clusters	50	55	60	65	70	75
Arousal	0.036	0.038	0.036	0.037	0.035	0.036
Expectancy	0.157	0.160	0.162	0.175	0.175	0.182
Power	0.033	0.032	0.033	0.032	0.031	0.032
Valence	0.050	0.048	0.048	0.047	0.047	0.045
Mean	0.069	0.069	0.069	0.072	0.072	0.074

usually spread out across different values rather than concentrated on one particular value range (See Figure VI). This could be caused by the complexity when interpreting human affective state. For example for the same facial expression an annotator could give different affect values in different situations. One way to address this problem is by finding the optimal cluster size to generate the best representative samples. As the details of how the cluster size is selected are not presented in the original paper, for our baseline system the optimal cluster size is selected as the largest cluster size with the lowest affect value variation in each cluster. Due to the size of training sample (around 8500 training samples after extracting the dynamic features), a large number of clusters could result in empty clusters. The cluster size is set to vary from 50 to 75 with a step of 5. The mean variation for different cluster size for each dimension for the dynamic shape features is shown in Table V. The optimal cluster size 60 is chosen since it gives lowest variance on arousal and valence while comparable variance on expectancy and power when compared to other cluster sizes. The same process is applied to the dynamic global appearance features, dynamic local appearance features and audio features. As a result 60 is selected as the cluster size for all the four features set and this in fact agrees with the optimal size selected by the original paper.

A.4.2 Optimising Kernel Width

As discussed in Section A.3.2.4, the kernel width σ is used as a smoothing factor to the kernel function. Again since the method of choosing the optimal value of σ is not detailed in the original paper, the most appropriate kernel width is selected based on the highest correlation score between the prediction and the labels. The kernel regression is trained on the training partition of the AVEC database and tested on the development partition.

A.4.3 Optimising Hyperparameters

In order to optimise the hyperparameters for the local linear regression in the fusion step, the baseline paper uses a subject-independent cross-validation on the training partition of the AVEC database. The training partition consists of 31 videos with 6 subjects. In our implementation a 6-fold cross-validation (data for 5 subjects are selected for training while the last one is used for testing) is used to select the best set of hyperparameters.

A.4.4 Results Comparison

The final results of the implementated system are shown in Table VI. As per the paper, the baseline system was trained on the training set and evaluated using the development set of the AVEC database. For each affect dimension, the left column shows the result reported in the paper, while the right column shows the result of the implemented baseline system. The results are measured in terms of Pearson's correlations averaged over all sequences.

The implementation generally agrees with the original paper i.e. that the shape features are effective at predicting the arousal and expectancy dimensions while the local appearance features are good at predicting valence and power dimensions. The local linear regression fusion method proposed by the baseline paper increases the

Table VI: Comparison between result reported in the paper and the implementation. Left: results reported in original paper. Right: results from our implementation

	Valence		Arousal		Expectancy		Power		Mean	
Shape	0.319	0.0630	0.538	0.1881	0.365	0.1571	0.429	0.168	0.413	0.1442
Global Appearance	0.281	0.0538	0.498	0.1754	0.347	0.1510	0.431	0.143	0.389	0.1267
Local Appearance	0.354	0.0752	0.470	0.1628	0.323	0.150	0.432	0.153	0.395	0.1293
Audio	-0.057	0.062	0.445	0.210	0.280	0.121	0.298	0.213	0.241	0.131
Fusion	0.350	0.082	0.644	0.236	0.341	0.187	0.511	0.213	0.461	0.168

overall performance of the prediction. However, due to lack of detailed description on various part of the baseline system, the implementation could not reproduce the high correlation scores reported by the paper. One of the main reasons could be the correlation-based feature selection part discussed in Section A.3.2.2. The baseline paper only focused on how the feature score is calculated, but there is no detailed information on how to select features based on the feature score. Additionally, the value of parameters of the system such as kernel width are also not given in the paper.

A.5 Conclusion

In this Appendix, the reason for choosing the proposed baseline system is described. Then the components of the baseline system are introduced along with the implementation details. Experiments carried out to optimise the parameters of the baseline system are then described. Although the implementation follows the general conclusion of the baseline system, there is a large difference between the reported results for the original paper and our implementation. Notwithstanding this, it is proposed that the implementation is sufficient for investigating the performance gain and various enhancements to an affect recognition system proposed in this thesis. Going forward, apart from the implemented winning system for AVEC 2012, the challenge baseline system for AVEC 2012 (Schuller *et al.*, 2012) and AVEC 2015 (Ringeval *et al.*, 2015b), as well as the winning system of for AVEC 2015 (He *et al.*,

2015) are also selected for comparison in the following chapters.

Appendix B

Head Nod and Shake Detection

B.1 Introduction

Nonverbal behaviours such as head gestures, body language, facial expression and eye contact play an import role in daily communications. As the most common head gestures, head nod and shake are usually used as semantic functions (e.g. nodding means yes, and shaking means no), affect indication (e.g. nodding means approval or acceptance) and conversational feedback (e.g. keep conversation flow), at least in Western Europe. Therefore, the detection of head nods and shakes can be seen as a valuable module for achieving affect recognition and natural human-computer interaction. In this paper we describe a new system that detects head nod and shake in real time. We use Microsoft Kinect and Kinect for Windows SDK to estimate head pose robustly. The direction of head movement is then determined based on the head pose and used by a discrete Hidden Markov Model (HMM) classifier as the observation sequence to detect whether head nod or shake occurs.

Much work has been done on head nod and shake detection stretching back over a decade. The related work presented by Davis and Vaks (2001) proposed a head gesture recognition system for interfaces. The IBM PupilCam is first used to obtain the location of the user's face. Based on the face location, a Timed Finite Sate

Machine is used to detect head nod and shake and the results are used to drive a perceptual dialogue-box agent (e.g., nod=YES). Similarly, Kapoor and Picard (2001) present a system that uses a customised IR camera for pupil tracking and a discrete Hidden Markov Model to detect head nod and shake. Kawato and Ohya (2000) described another system to detect head nods and shakes in real time by directly detecting and tracking the between-eyes region using a webcam. Combining the circle frequency filter together with skin colour information and template, the between-eyes region could be detected and tracked. A rule based detection algorithm is then applied to the movement of the between-eyes region to detect head nods and shakes. Because of its simple rule based detection, some non-regular head nods and shakes may not be detected. It should be noted that all the systems mentioned above need to track the eye pupil position in order to detect head nod and shake, and will not be able to detect any head gesture if the user's eyes are closed. The work carried out by Tan and Rong (2003) present another method to detect head nodding and shaking in real time from video streams. The AdaBoost algorithm is first used to detect the user's face and based on the physiological information of the eye location in the face, eye location can be obtained in each frame. The direction of head movement is calculated based on eye location and used as an observation sequence for a discrete HMM to detect head nods and shakes. Kwon *et al.* (2006) present a new method for head nods and shakes detection by using 3D cylindrical head model (CHM) and dynamic template to estimate head pose and use the accumulative Hidden Markov Models to detect head nod and shake.

In this section, we present a new method to robustly detect head nods and shakes in real time using the Microsoft Kinect. Despite a lot of work on head nod/shake in the past, to the best of our knowledge this has not been widely explored with the Kinect due to its relatively recent introduction. We first use the Microsoft Kinect for Windows SDK to estimate the head pose of the user. The change of head pose in each frame indicates the direction of the head movements that are then used as an observation sequence by a discrete Hidden Markov Models (HMMs) to detect if a head nod or shake occurs. The proposed system runs fast and can detect the head nods and shakes in real time on a standard desktop PC. The approach can also robustly detect non-obvious and non-regular head nods and shakes.

The overall architecture of the system is shown in Figure. I. The head pose is first obtained from Kinect through the Kinect for Windows SDK. Then, the head pose in a temporal window is analysed as a sequence of head movements. Finally, we use three HMMs to detect the presence of head shake, head nod and other head gestures in this sequence of head movements. The largest likelihood value is selected as the detection result. In order to further distinguish head nod and shake from other head gestures, a predefined threshold is used. More details are described in this section.

B.2 Head Pose Estimation

Head pose estimation has received a lot of attention recently as a key element of human behaviour analysis. With depth cameras such as Microsoft Kinect becoming available at commodity prices, the research focus of head pose estimation have shift from 2D video data based to depth data based and have shown very good results compared to 2D approach (Breitenstein *et al.*, 2008; Fanelli *et al.*, 2011). The Microsoft Kinect supports the capture of 2D RGB streams and 3D depth streams at 30 frames per second, based on infrared projection and light coding techniques. However, the Kinect depth information is not very accurate and much noisier compared to the data obtained from other devices, such as a laser-scanner, for example. In order to estimate the head pose, the method described by Cai *et al.* (2010) was used. The method utilises a regularised maximum likelihood deformable model fitting (DMF) algorithm to reduce the effect of the noisy depth map acquired from the Kinect and to improve the accuracy of the estimation results. As this method has been implemented in the recent release of Kinect for Windows SDK, we use it

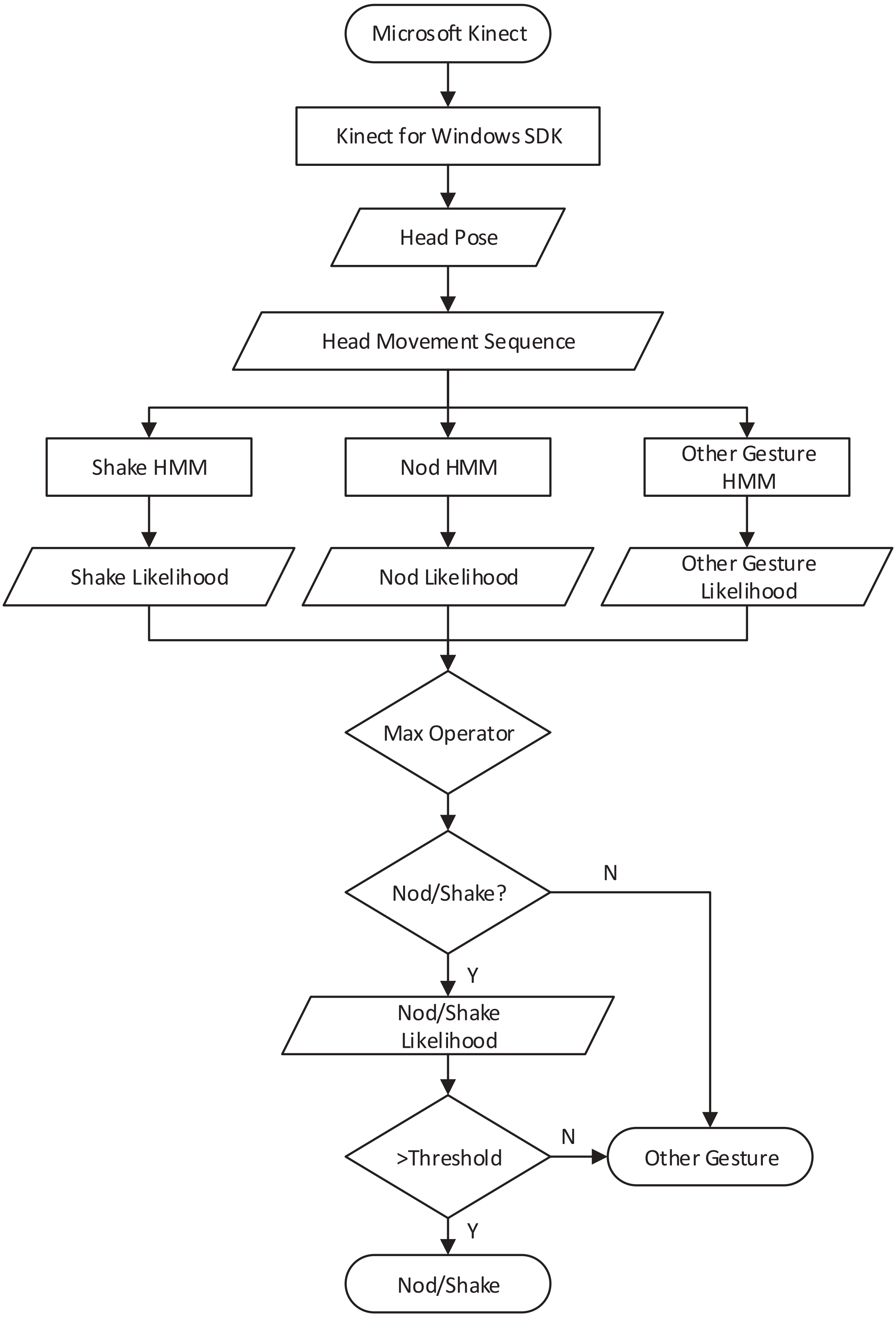


Figure I: System Overview

directly to obtain the head pose of the user. The SDK gives head pose with respect to the Kinect by three angles: pitch, roll and yaw, as illustrated in Figure. II. The angles are expressed in degrees, with values ranging from -90 to +90 degrees.

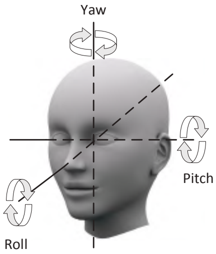
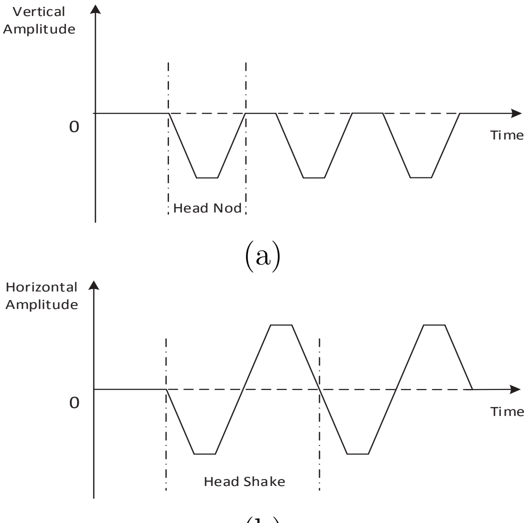


Figure II: Yaw, Pitch and Roll



(a)

(b)

Figure III: (a) Typical Nod Sequence. (b) Typical Shake Sequence.

B.3 Head Nod and Shake Detection

Although head nods and shakes could be performed differently by different people in terms of intervals and amplitudes, some common characteristics still exist for the head movement to be recognized as nods or shakes. In this paper we consider a nod as the head tilted in an alternating down and up manner, whereas a shake is rotation of the head horizontally from side-to-side. This is shown in Figure. III. The vertical and horizontal movement could be represented by pitch and yaw in terms of head pose as shown in Figure. II. By comparing the difference of pitch and yaw in two adjacent frames, the direction of head movement can then be determined. Following Kapoor and Picard (2001), the direction is represented by five directional symbols (Up, Down, Left, Right and Still). Based on the five states of head movements, three Hidden Markov Models (HMM) termed nodHMM, shakeHMM and otherHMM are trained. The nodHMM consists of three states Up, Down and Still, whereas the shakeHMM consists Left, Right and Still. Both HMMs have five observation states Up, Down, Left, Right and Still. To further distinguish other head movements (E.g. moving up, moving down, moving left and moving right) from the actual head nods and shakes, we first build an additional HMM, termed otherHMM, which consists of five states Up, Down, Left, Right and Still to recognize head gestures except head nods and shakes, and then we compare the nod or shake likelihood values to a predefined threshold. The state transitions of head nod, head shake and other gestures is shown in Figure. IV.

In order to analyse head movement continuously, we choose a window size of 0.6 seconds similar to the work by Tan and Rong (2003), corresponding to 18 frames/sec, which we found sufficient to detect both slow as well as subtle head nods and shakes. During the training phase, we extract the head pose using the method mentioned in section B.2 for each frame and formed an observation sequence of 18 frames. Since it is impossible for 18 frames to compromise all the actions of head nod and shake we consider the sequence containing down as head nod and any obvious Left or Right as head shake too. The Baum Welch algorithm (Rabiner, 1989) is used to train the nodHMM, shakeHMM and otherHMM based on the observation sequence. In the testing phase, the forward-backward procedure (Rabiner, 1989) is used to compute the log likelihood for the input observation sequence on three HMMs. The largest likelihood value is selected, and if a head nod or shake is detected, it is further compared to a predefined threshold. If the likelihood value is larger than

	Recognized As		
	Head Nods	Head Shakes	Other
Head Nods	22	0	3
Head Shakes	0	23	2
Other	4	1	20

Table I: Recognition Results for Training set

Table II: Recognition Results for Testing set

	Recognized As		
	Head Nods	Head Shakes	Other
Head Nods	21	0	4
Head Shakes	0	22	3
Other	5	1	19

the predefined threshold the observation sequence is considered to be a nod or a shake, otherwise it is considered to be other head gesture such as still or looking upward.

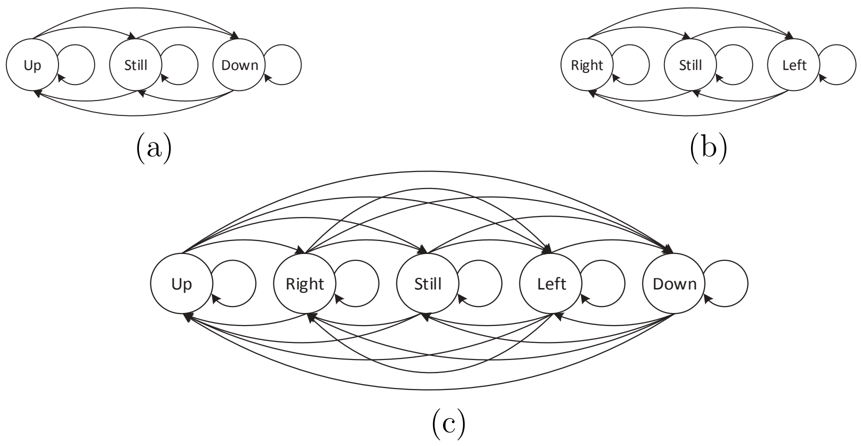


Figure IV: (a) Transition of nodHMM's hidden states. (b) Transition of shakeHMM's hidden states. (c) Transition of otherHMM's hidden states

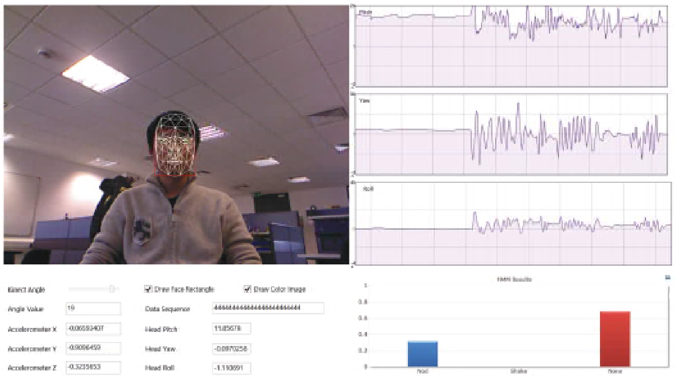
B.4 Recognition Results

We collected a database of manually performed head nods, shakes and other gestures to train the HMMs. Microsoft Kinect and Kinect studio were used to capture the head motion. In total 150 samples with 50 head nods, 50 head shakes and 50 other head gestures (including still, look upward, look downward, look leftward and look rightward) were collected and manually annotated. These head nods and shakes are of obvious motions of nod and shake in different motion magnitudes. Thus, we ensured that the trained HMM classifiers were suitable for the head nods or shakes with small or big head motions. A random 50% of the each gesture class is selected for training to estimate the parameters of nod, shake and other HMMs.

After training, the estimated parameters and the detection algorithm were implemented on an Intel Core i7 3.4GHz machine with Windows 7 with the Kinect placed under the monitor. The details of recognition results are shown in Table I and II. This performance appears to be comparable, if not better, to the results obtained by other methods, such as in the work by Kapoor and Picard (2001) and Tan and Rong (2003). Future work will investigate this more fully by applying those techniques to our dataset.

From the results we can see there is no mis-classification among head nods and head shakes. Most missed head nods are due to the head gestures such as look downward and look upward and missed head shakes are due to look leftward and look rightward motions.

A demonstration system has been developed to visualise the estimated head pose value and show the detection results in a bar chart form, shown in Figure. V. When the head is detected by the Kinect, a 3D mesh will appear on the face. At the same time, the detection of head nod and shake begins to work. We visualise the Pitch, Yaw and Roll data from Kinect for Windows SDK. The real-time data of Pitch, Yaw, Raw, number of sequence, and the head position relative to the position of the Kinect is displayed. We finally show the detection results of Nod, Shake and None by HMM classifiers with above data. The classifier with the maximum probability is the final detection results.



☑ Draw Face Rectangle ☑ Draw Color Image

Kinect Angle

Angle Value: 19

Data Sequence: 4444444444444444444444

Accelerometer X: -0.06593407

Head Pitch: 11.85678

Accelerometer Y: -0.9006459

Head Yaw: -0.0970258

Accelerometer Z: -0.3235853

Head Roll: -1.110691

Figure V: Screenshot of the system in operation

Bibliography

S. Afzal and P. Robinson. Natural affect datacollection & annotation in a learning context. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009.

T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 356–361. IEEE, 2013.

N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

M. B. Arnold. Emotion and personality. 1960.

A. F. Ax. The physiological differentiation between fear and anger in humans. *Psychosomatic medicine*, 15(5):433–442, 1953.

D. L. Baggio. *Mastering OpenCV with practical computer vision projects.* Packt Publishing Ltd, 2012.

T. Baltrusaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson, and R. Picard. Real-time inference of mental states from facial expressions and upper body gestures. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 909–914. IEEE, 2011.

T. Baltrusaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

S. Baron-Cohen, O. Golan, S. Wheelwright, and J. Hill. Mind reading: The interactive guide to emotions. *London: Jessica Kingsley Ltd Google Scholar*, 2004.

M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573. IEEE, 2005.

A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong. " you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *LREC*, 2004.

S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo. 3d facial expression recognition using sift descriptors of automatically detected keypoints. *The Visual Computer*, 27(11):1021–1036, 2011.

M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister. Real-time face pose estimation from single range images. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

H. Brugman and A. Russel. Annotating multi-media/multi-modal resources with elan. 2004.

S. Burger, V. MacLaren, and H. Yu. The isl meeting corpus: the impact of meeting type on speech style. 2002.

W. Burgin, C. Pantofaru, and W. D. Smart. Using depth information to improve face detection. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 119–120. ACM, 2011.

Q. Cai, D. Gallup, C. Zhang, and Z. Zhang. 3d deformable face tracking with a commodity depth camera. In *Computer Vision–ECCV 2010*, pages 229–242. Springer, 2010.

R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas. *The Oxford handbook of affective computing*. Oxford University Press, USA, 2014.

A. Camurri, B. Mazzarino, and G. Volpe. Analysis of expressive gesture: The eyesweb expressive gesture processing library. In *Gesture-based communication in human-computer interaction*, pages 460–467. Springer, 2004.

G. Caridakis, K. Karpouzis, and S. Kollias. User and context adaptive neural networks for emotion recognition. *Neurocomputing*, 71(13):2553–2562, 2008.

G. Chanel, K. Ansari-Asl, and T. Pun. Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In *2007 IEEE International Conference on Systems, Man and Cybernetics*, pages 2662–2667. IEEE, 2007.

L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2015.

L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu. Multimodal human emotion/expression recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 366–371. IEEE, 1998.

L. S.-H. Chen. *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction*. PhD thesis, Citeseer, 2000.

S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56. ACM, 2015.

CMV, University of OULU. URL `http://www.cse.oulu.fi/CMV/Downloads/`.

J. F. Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238. ACM, 2006.

J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(02):121–132, 2004.

A. Colombo, C. Cusano, and R. Schettini. 3d face detection using curvature analysis. *Pattern recognition*, 39(3):444–455, 2006.

J. Cong and B. Xiao. Minimizing computation in convolutional neural networks. In *International Conference on Artificial Neural Networks*, pages 281–290. Springer, 2014.

T. Cootes. An introduction to active shape models.

T. F. Cootes, C. J. Taylor, *et al.* Statistical models of appearance for computer vision.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

M. Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2):117–139, 2004.

R. Cowie and M. Sawey. Gtrace-general trace program from queens, belfast. 2011.

R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

R. Cowie, E. Douglas-Cowie, J.-C. Martin, and L. Devillers. *The essential role of human databases for learning in and validation of affectively competent agents.* Oxford: OUP, 2010a.

R. Cowie, H. Gunes, G. McKeown, L. Vaclau-Schneider, J. Armstrong, and E. Douglas-Cowie. The emotional and communicative significance of head nods and shakes in a naturalistic database. In *Proc. of LREC Int. Workshop on Emotion*, pages 42–46, 2010b.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

C. Darwin. 1965. the expression of the emotions in man and animals. *London, UK: John Marry*, 1872.

C. Darwin. *The expression of the emotions in man and animals.* Oxford University Press, 1998.

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, 2(7): 1160–1169, 1985.

J. W. Davis and S. Vaks. A perceptual user interface for recognizing head gesture acknowledgements. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–7. ACM, 2001.

F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1970–1973. IEEE, 1996.

L. Devillers and I. Vasilescu. Reliability of lexical and prosodic cues in two real-life spoken dialog corpora. In *LREC*, 2004.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, *et al.* The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In *International Conference on Affective Computing and Intelligent Interaction*, pages 488–500. Springer, 2007.

I. L. Dryden and K. V. Mardia. *Statistical shape analysis*, volume 4. Wiley Chichester, 1998.

G. Dunn. Morphometric tools for landmark data: Geometry and biology. *Statistics in Medicine*, 12(7):714–715, 1993. ISSN 1097-0258.

P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation.* University of Nebraska Press, 1971.

P. Ekman and W. V. Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

P. Ekman and W. V. Friesen. Facial action coding system: A technique for the measurement of facial movement. palo alto, 1978.

P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, USA, 1997.

P. Ekman, R. W. Levenson, and W. V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210, 1983.

M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

H. P. Espinosa, C. A. R. García, and L. V. Pineda. Features selection for primitives estimation on emotional speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5138–5141. IEEE, 2010.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008.

G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010.

G. Fanelli, T. Weise, J. Gall, and L. Van Gool. Real time head pose estimation from consumer depth cameras. In *Pattern Recognition*, pages 101–110. Springer, 2011.

J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

E. Fox. *Emotion science cognitive and neuroscientific approaches to understanding human emotions*. Palgrave Macmillan, 2008.

N. Fragopanagos and J. G. Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.

N. H. Frijda and J. Swagerman. Can computers feel? theory and design of an emotional system. *Cognition and emotion*, 1(3):235–257, 1987.

D. George and P. Mallery. Spss for windows step by step: A simple guide and reference 11.0 update, 2003.

D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer. Technique for automatic emotion recognition by body gesture analysis. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.

B. Gong, Y. Wang, J. Liu, and X. Tang. Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 569–572. ACM, 2009.

M. Grimm and K. Kroschel. *Emotion estimation in speech using a 3d emotion space concept*. na, 2005.

M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pages 865–868. IEEE, 2008.

T. Gritti, C. Shan, V. Jeanne, and R. Braspenning. Local features based facial expression recognition with face registration errors. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–8. IEEE, 2008.

H. Gunes. Automatic, dimensional and continuous emotion recognition. 2010.

H. Gunes and M. Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Intelligent virtual agents*, pages 371–377. Springer, 2010.

H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1148–1153. IEEE, 2006.

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

H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2): 120–136, 2013.

N. Halko, P.-G. Martinsson, Y. Shkolnisky, and M. Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific computing*, 33(5):2580–2594, 2011.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM, 2015.

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

K.-C. Huang, S.-Y. Huang, and Y.-H. Kuo. Emotion recognition based on a novel triangular facial feature extraction method. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–6. IEEE, 2010.

G. L. Huttar. Relations between prosodic variables and emotions in normal american english utterances. *Journal of Speech, Language, and Hearing Research*, 11(3): 481–487, 1968.

C. E. Izard. The face of emotion. 1971.

R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

W. James. Ii.what is an emotion? *Mind*, (34):188–205, 1884.

Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 36(5):862–875, 2006.

B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE, 2011.

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

X. Jin and X. Tan. Face alignment in-the-wild: A survey. *arXiv preprint arXiv:1608.04188*, 2016.

M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels. Ensemble methods for continuous affect recognition: multi-modality, temporality, and challenges. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 9–16. ACM, 2015.

T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

A. Kapoor and R. W. Picard. A real-time head nod and shake detector. In *Pro-*

ceedings of the 2001 workshop on Perceptive user interfaces, pages 1–5. ACM, 2001.

K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Artifical intelligence for human computing*, pages 91–112. Springer, 2007.

F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab. Human-computer interaction: Overview on state of the art. 2008.

S. Kawato and J. Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the between-eyes. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 40–45. IEEE, 2000.

H. Kaya, F. Çilli, and A. A. Salah. Ensemble cca for continuous emotion prediction. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 19–26. ACM, 2014.

P. Khorrami, T. L. Paine, K. Brady, C. Dagli, and T. S. Huang. How deep neural networks can improve emotion recognition on video data. *arXiv preprint arXiv:1602.07377*, 2016.

D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

D. E. King. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015.

M. Kipp. Anvil-a generic annotation tool for multimodal dialogue" in procs of 7th european conference on speech communication and technology. 2001.

M. Kipp. ANVIL 5.0: Overview of new features. `http://embots.dfki.de/anvil/forum/viewtopic.php?f=5&t=19`, 2010. [Online; accessed 29-May-2014].

H. Kobayashi and F. Hara. The recognition of basic facial expressions by neural network. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pages 460–466. IEEE, 1991.

S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

O. Kwon, J. Chun, and P. Park. Cylindrical model-based head tracking and 3d pose recovery from sequential face images. In *Proceedings of the 2006 International Conference on Hybrid Information Technology-Volume 01*, pages 135–139. IEEE Computer Society, 2006.

P. J. Lang. The emotion probe: Studies of motivation and attention. *American psychologist*, 50(5):372, 1995.

I. Lawrence and K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

R. S. Lazarus. Psychological stress and the coping process. 1966.

A. Lee. virtualdub.org, 2013. URL `http://www.virtualdub.org/`.

C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, 13(2):293–303, 2005.

R. W. Levenson. Autonomic nervous system differences among emotions. *Psychological science*, 3(1):23–27, 1992.

B. Y. Li, A. S. Mian, W. Liu, and A. Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 186–192. IEEE, 2013.

X. Li, Q. Ruan, and Y. Ming. 3d facial expression recognition based on basic geometric features. In *IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS*, pages 1366–1369. IEEE, 2010.

P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and K. M. Prkachin. Automatically detecting pain using facial actions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–8. IEEE, 2009.

P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

S. Lucey, A. B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through aam representations of the face. *Face recognition*, pages 275–286, 2007.

A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti. Local 3d shape analysis for facial expression recognition. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4129–4132. IEEE, 2010.

M. Mahmoud, T. Baltrušaitis, P. Robinson, and L. D. Riek. 3d corpus of spontaneous complex mental states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 205–214. Springer, 2011.

M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 336–342. IEEE, 2011.

M. K. Mandal, A. Awasthi, *et al. Understanding Facial Expressions in Communication*. Springer, 2015.

S. Marčelja. Mathematical description of the responses of simple cortical cells*. *JOSA*, 70(11):1297–1300, 1980.

R. Mattheij, E. Postma, Y. Van den Hurk, and P. Spronck. Depth-based detection using haarlike features. 2012.

I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.

A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 1995.

A. Metallinou, A. Katsamanis, and S. Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152, 2013.

A. Mian, M. Bennamoun, and R. Owens. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 29(11), 2007.

Michel Valstar. URL http://www.cs.nott.ac.uk/~pszmv/resources/visum_package.zip.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

P. Nair and A. Cavallaro. 3-d face detection, landmark localization, and registration using a point distribution model. *IEEE Transactions on multimedia*, 11(4):611–623, 2009.

A. Nakasone, H. Prendinger, and M. Ishizuka. Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th International Workshop on Biosignal Interpretation*, pages 219–222. Citeseer, 2005.

H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347. IEEE, 2014.

M. A. Nicolaou. Discrete & continuous audio-visual recognition of spontaneous emotions. Master's thesis, 2009.

M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

M. A. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic probabilistic cca for analysis of affective behaviour. In *European Conference on Computer Vision*, pages 98–111. Springer, 2012.

J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 501–508. ACM, 2012.

T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision &amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 582–585. IEEE, 1994.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *Image and signal processing*, pages 236–243. Springer, 2008.

A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

A. Ortony, G. L. Clore, and A. Collins. *The cognitive structure of emotions.* Cambridge university press, 1988.

C. E. Osgood, W. H. May, and M. S. Miron. *Cross-cultural universals of affective meaning.* University of Illinois Press, 1975.

A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, 2005.

D. Ozkan, S. Scherer, and L.-P. Morency. Step-wise emotion recognition using concatenated-hmm. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 477–484. ACM, 2012.

M. Pantic and M. S. Bartlett. Machine analysis of facial expressions. 2007.

M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005.

C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Computer vision, 1998. sixth international conference on*, pages 555–562. IEEE, 1998.

R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(10):1175–1191, 2001.

J. Platt *et al.* Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

S. Ramanathan, A. Kassim, Y. Venkatesh, and W. S. Wah. Human facial expression recognition using a 3d morphable model. In *2006 International Conference on Image Processing*, pages 661–664. IEEE, 2006.

F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015a.

F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. The av+ec 2015 multimodal affect recognition challenge: Bridging across audio, video, and physiological data. 2015b.

M. Rosato, X. Chen, and L. Yin. Automatic registration of vertex correspondences for 3d facial expression analysis. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–7. IEEE, 2008.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

J. Russel. A circumplex model of affect. *Journal ofPersonalityand*, 1980.

J. A. Russell and L. F. Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999.

E. Sánchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, and J. L. Alba-Castro. Audiovisual three-level fusion for continuous estimation of russell's emotion circumplex. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 31–40. ACM, 2013.

G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008.

A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 485–492. ACM, 2012a.

A. Savran, B. Sankur, and M. T. Bilge. Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. *Pattern recognition*, 45(2):767–782, 2012b.

K. R. Scherer and J. S. Oshinsky. Cue utilization in emotion attribution from auditory stimuli. *Motivation and emotion*, 1(4):331–346, 1977.

K. R. Scherer, T. Bänziger, and E. Roesch. *A Blueprint for Affective Computing: A sourcebook and manual.* Oxford University Press, 2010.

K. L. Schmidt and J. F. Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American journal of physical anthropology*, 116(S33):3–24, 2001.

B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012.

J. Segal. *The Language of Emotional Intelligence: The Five Essential Tools for Building Powerful and Effective Relationships.* McGraw Hill professional. McGraw-Hill Education, 2008. ISBN 9780071544566. URL `https://books.google.ie/books?id=w0efbfeiIfcC`.

T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4): 993–1005, 2012.

M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk. Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 57–63. ACM, 2014.

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

J. Shi, A. Samal, and D. Marx. How effective are landmarks and their geometry for face recognition? *Computer vision and image understanding*, 102(2):117–133, 2006.

A. Smola and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

H. Soyel and H. Demirel. Facial expression recognition using 3d facial feature distances. In *International Conference Image Analysis and Recognition*, pages 831–838. Springer, 2007.

M. Stewart, C. Gwen, G. Mark, R. Ian, R. Javier, *et al.* Automatic recognition of facial actions in spontaneous expressions. 2006.

G. Stratou, A. Ghosh, P. Debevec, and L.-P. Morency. Exploring the effect of illumination on automatic expression recognition using the ict-3drfe database. *Image and Vision Computing*, 30(10):728–737, 2012.

M. Suwa, N. Sugie, and K. Fujimora. A preliminary note on pattern recognition of human emotional expression. In *International joint conference on pattern recognition*, pages 408–410, 1978.

K. Takahashi. Remarks on emotion recognition from multi-modal bio-potential signals. In *Industrial Technology, 2004. IEEE ICIT'04. 2004 IEEE International Conference on*, volume 3, pages 1138–1143. IEEE, 2004.

TalkingFaceVideo. Talking face video @ONLINE, 2002. URL `http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html`.

W. Tan and G. Rong. A real-time head nod and shake detector using hmms. *Expert Systems with Applications*, 25(3):461–466, 2003.

H. Tang and T. S. Huang. 3d facial expression recognition based on properties of line segments connecting facial feature points. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

S. Tomkins. Affect/imagery/consciousness. vol. 2: The negative affects. 1963.

S. S. Tomkins. Affect, imagery, consciousness: Vol. i. the positive affects. 1962.

J. Trouvain and W. J. Barry. The prosody of excitement in horse race commentaries. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

F. Tsalakanidou and S. Malassiotis. Robust facial action recognition from real-time 3d streams. In *Computer Vision and Pattern Recognition Workshops*, pages 4–11, 2009.

P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. Frank, and P. Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 71(2):197–214, 2007.

M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. Intl Conf. Language Resources and Evaluation, Wshop on EMOTION*, pages 65–70, 2010.

M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

M. F. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(1):28–43, 2012.

M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 38–45. ACM, 2007.

J. Van den Stock, R. Righart, and B. De Gelder. Body expressions influence recognition of emotions in the face and voice. *Emotion*, 7(3):487, 2007.

L. Van Der Maaten. Audio-visual emotion challenge 2012: a simple approach. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 473–476. ACM, 2012.

V. Vapnik and A. Y. Chervonenkis. Algorithms with complete memory and recurrent algorithms in the problem of learning pattern recognition. *Avtomat. i Telemekh*, (4):95–106, 1968.

V. Vapnik and A. Y. Chervonenkis. Theory of uniform convergence of frequencie of appearance of attributes to their probabilities and problems of defining optimal solution by empiric data. *Avtomatika i Telemekhanika*, 2:42–53, 1971.

V. N. Vapnik. The nature of statistical learning theory. 1995.

A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–4. IEEE, 2009.

P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

N. Vretos, N. Nikolaidis, and I. Pitas. 3d facial expression recognition using zernike moments on depth images. In *2011 18th IEEE International Conference on Image Processing*, pages 773–776. IEEE, 2011.

H. G. Wallbott. Bodily expression of emotion. *European journal of social psychology*, 28(6):879–896, 1998.

J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1399–1406. IEEE, 2006.

N. Wang, X. Gao, D. Tao, and X. Li. Facial feature point detection: A comprehensive survey. *arXiv preprint arXiv:1410.1037*, 2014.

J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie, and Z. Fu. Multimodal continuous affect recognition based on lstm and multiple kernel learning. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2014.

F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer. On the acoustics of emotion in audio: what speech, music, and sound have in common. 2013.

F. Weninger, J. Bergmann, and B. W. Schuller. Introducing currennt: the munich open-source cuda recurrent neural network toolkit. *Journal of Machine Learning Research*, 16(3):547–551, 2015.

C. Whissell. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4(113-131):94, 1989.

C. Whissell. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language 1, 2. *Psychological reports*, 105(2):509–521, 2009.

A. Wierzbicka. Talking about emotions: Semantics, culture, and cognition. *Cognition & Emotion*, 6(3-4):285–319, 1992.

C. E. Williams and K. N. Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, 1972.

M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *INTERSPEECH*, volume 2008, pages 597–600, 2008.

M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *INTERSPEECH*, pages 2362–2365, 2010.

W. Wundt. Fundamentals of psychology. *Liepzig. Engelman*, 1905.

X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.

L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.

L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pages 1–6. IEEE, 2008.

S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 2015.

Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical report, 2010.

X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum. Finding celebrities in billions of web images. *IEEE Transactions on Multimedia*, 14(4):995–1007, 2012.

X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 666–673. Ieee, 1999.

G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

Q. Zhen, D. Huang, Y. Wang, and L. Chen. Lpq based static and dynamic modeling of facial expressions in 3d videos. In *Biometric Recognition*, pages 122–129. Springer, 2013.