RECOGNISING COMPLEX MENTAL STATES FROM NATURALISTIC HUMAN-COMPUTER INTERACTIONS

By

Hamed Monkaresi

MSc., BSc.

A thesis submitted in fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Faculty of Engineering and Information Technologies

University of Sydney

March 2014

Thesis Supervisor: A/Prof Rafael A. Calvo

Associate Supervisor: Prof. Hong Yan

School of Electrical and Information Engineering,

The University of Sydney, Australia

DEDICATION

To my wife and family.

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my supervisor A/Professor Rafael Calvo, for giving me the opportunity to be part of this inspirational environment and for being supportive during my research. I would also like to thank my associate supervisor Professor Hong Yan for his valuable advice and comments.

I would also thank the Iranian Ministry of Science, Research, and Technology for supporting me with the primary scholarship during my candidature.

My sincere thanks also goes to Professor Peter Robinson, for inviting me to visit his group (Rainbow Group) at the University of Cambridge, and for his valuable advises on my project. I would also thank Dr. Sidney D'Mello for his critical evaluation of my papers and his valuable suggestions.

I would especially like to thank the former and current members of our research group Learning & Affect Technologies (LATTE) especially Sazzad Hussain, Payam Aghaei Pour, and Omar Alzoubi for their support, cooperation, and kindness.

A special thanks to my family. Words cannot express how grateful I am to my mother, and father for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. At the end I would like express appreciation to my beloved wife Mahya who was always my support in the moments when I was really needed.

PREFACE

The following is a list of publications in peer-reviewed journals and conference proceedings produced during my Ph.D. candidature:

1. Monkaresi, H., Bosch N., Calvo, R. A., D'Mello, S. K., & Robinson P., (To be submitted). Detecting Task Engagement using Facial Expression and Heart Rate. *IEEE Transactions on Affective Computing*.

2. Monkaresi, H., Calvo, R. A., & Hussain, M. S. (2014). Using Remote Heart Rate Measurement for Affect Detection. In *The 27th International FLAIRS Conference* (pp. 118–123). Pensacola Beach, Florida, USA: AAAI.

3. Monkaresi, H., Calvo, R. A., & Yan, H. (2014). A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam. *IEEE Journal of Biomedical and Health Informatics*, *18*(4), 1153–1160. doi:10.1109/JBHI.2013.2291900

4. Monkaresi, H., Hussain, M. S., & Calvo, R. A. (2012). A Dynamic Approach for Detecting Naturalistic Affective States from Facial Videos during HCI. In M. Thielscher & D. Zhang (Eds.), *AI 2012: Advances in Artificial Intelligence* (pp. 170–181). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-35101-3_15

5. Monkaresi, H., Hussain, M. S., & Calvo, R. A. (2012). Classification of Affects Using Head Movement, Skin Color Features and Physiological Signals. In *IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2012)* (pp. 2664–2669). doi:10.1109/ICSMC.2012.6378149

6. Monkaresi, H., Calvo, R. A., & Hussain, M. S. (2012). Automatic natural expression recognition using head movement and skin color features. In Genny Tortora, Stefano Levialdi, & Maurizio Tucci (Eds.), *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)* (pp. 657–660). New York: ACM. doi:10.1145/2254556.2254678

7. Hussain, M. S., Monkaresi, H., & Calvo, R. A. (2012). Categorical vs. Dimensional Representations in Multimodal Affect Detection during Learning. In *11th International Conference on Intelligent Tutoring Systems (ITS '12)* (pp. 78–83). Berlin / Heidelberg: Springer. doi:10.1007/978-3-642-30950-2_11

8. Hussain, M. S., Monkaresi, H., & Calvo, R. A. (2012). Combining Classifiers in Multimodal Affect Detection. In *Proceedings of the Tenth Australasian Data Mining Conference (AusDM 2012)* (pp. 103–108). Sydney, Australia.

Chapter 3 of this thesis includes all the methods reported in papers 1 to 5. The thesis author was the primary investigator and author of papers 1, 2, and 3. In papers 4 and 5, other investigators did the data collections but the main author undertook the data processing and analysis.

Chapter 4 of the thesis is a reprint of the results section of the third paper. This paper has been accepted by the *IEEE Journal of Biomedical and Health Informatics.* The main investigator wrote most parts of the paper.

Chapter 5 of the thesis is based on the results reported in papers 2, 4, and 5. The primary author wrote most parts of the second paper, including data collection, data processing, and analysis. The first author wrote the data analysis and discussion parts of the fourth and fifth papers.

Chapter 6 of this thesis is revised and going to be submitted to the *IEEE Transactions on Affective Computing* (paper 1).

The seventh and eighth papers have contributed partially to this thesis. The first investigator wrote these two papers. The second author did the video processing and extracted the facial features.

ABSTRACT

New advances in computer vision techniques will revolutionize the way we interact with computers, as they, together with other improvements, will help us build machines that understand us better. Combining depth sensors with video data can improve the accuracy of recognizing facial expression, body gestures, and posture. The face is the main non-verbal channel for human-human communication and contains valuable information about emotion, mood, and mental state. Affective computing researchers have investigated widely how facial expressions can be used for automatically recognizing affect and mental states. Nowadays, physiological signals can be measured by video-based techniques, which can be utilised for emotion detection. Physiological signals, are an important indicator of internal feelings and affective states, and are more robust against social masking compared with other modalities, such as facial expression and body gesture.

This thesis focuses on computer vision techniques to detect facial expression and physiological changes for recognizing non-basic and natural emotions during human-computer interaction. It covers all stages of the research process from data acquisition, integration and application.

Most previous studies focused on acquiring data from prototypic basic emotions acted out under laboratory conditions. To evaluate the proposed method under more practical conditions, two different scenarios were used for data collection. In the first scenario, a set of controlled stimulus was used to trigger the user's emotion. The second scenario aimed at capturing more naturalistic emotions that might occur during a writing activity. In the second

scenario, the engagement level of the participants with other affective states was the target of the system.

For the first time this thesis explores how video-based physiological measures can be used in affect detection. Video-based measuring of physiological signals is a new technique that needs more improvement to be used in practical applications. A machine learning approach is proposed and evaluated to improve the accuracy of heart rate (HR) measurement using an ordinary camera during a naturalistic interaction with computer. Extracted information from measured HRs is used for affect detection.

We evaluated the integration of data from multiple modalities, a popular method that has received more attention recently in the affect detection research area. It has been shown that adding more channels and modalities can improve the performance of the affect detection systems. We have combined three sorts of features that can be extracted from video modality: HR features, and appearance-based and geometric-based features from facial expression and head movements.

Overall, the results indicate that affect detection is more accurate under controlled conditions compared with the naturalistic situations. In the first scenario (controlled stimulus), fusing the HR features with the appearance-based and geometric-based features obtained the best results for affect detection using user-dependent models. The fusion model also showed a significant improvement over individual channels for detecting normative affective labels in the user-independent analysis. However, the improvements of the fusion model were not significant for affect detection in the naturalistic scenario. For engagement detection, the fusion model outperformed individual channels in the user-dependent analysis during the naturalistic interactions. Again, the fusion approach did not improve the accuracy of engagement detection using the user-independent models. The results also showed that building gender-

specific models improved the accuracy of affect detection slightly for each channel compared with the general model.

Analysing single channel showed interesting results. Although the HR channel achieved poor Kappa scores for detecting valence and arousal, the positive values of the Kappa scores showed the possibility of using the HR channel for affect detection. The HR channel obtained a moderate accuracy for detecting engagement during naturalistic interaction. It should be mentioned that these HR features were extracted by a video-based method, which is improved to be adopted in the practical application. Our improvement decreased the root mean squared error from 43.76 beats per minute (bpm) to 3.64 (bpm).

We also realised that the eye region was more informative for engagement detection than the mouth region. This is also true for detecting valence and arousal during naturalistic interaction. Our finding also suggested that both appearance-based and motion-based features are essential for affect detection. The significant impact of distractions on the engagement level during naturalistic interaction has also been reflected in our findings.

ACRONYMS

AC	Affective Computing
ANOVA	Analysis Of Variance
ANS	Autonomic Nervous System
ANU	Animation Unit
AU	Action Unit
AVC	Average Vote Classifier
BLSTM	Bidirectional Long Short-Term Memory
BSS	Blind Source Separation
BVP	Blood Volume Pressure
CFS	Correlation Based Feature Selection
Chi2	Chi-Square
DBN	Dynamic Bayesian Network
DFA	Discriminant Function Analysis
DT	Dynamic Texture
ECG	Electrocardiogram
EEG	Electroencephalography
EMG	Electromyogram
ESI	Emergency Severity Index
F1	F-Measure
FACS	Facial Action Coding System
FER	Facial Expression Recognition
FFD	Free-form Deformations
FN	False Negative
FP	False Positive
fps	Frames Per Seconds
FT	Face Tracker
GSR	Galvanic Skin Response
HCI	Human Computer Interaction
HMM	Hidden Markov Model
HR	Heart Rate
HRV	Heart Rate Variability
IAPS	International Affective Picture System

ICA	Independent Component Analysis
ITS	Intelligent Tutoring Systems
J48	Decision Trees
JADE	Joint Approximate Diagonalization of Eigenmatrices
KNN	K-Nearest Neighbour
LBP	Local Binary Pattern
LBPTOP	Local Binary Pattern in Three Orthogonal Planes
LoA	Limits of Agreement
MAUI	Multimodal Affective User Interface
MHI	Motion History Images
MPA	Maximum Peak among All
MSE	Mean Absolute Error
NN	Neural Network
OpenCV	Open Computer Vision Library
PNS	Parasympathetic Nervous System
PPG	Photoplethysmography
PSD	Power Spectrum Density
QD	Quadratic Programming
RMSE	Root Mean Squared Error
ROI	Region Of Interest
RSP	Respiration
SAL	Sensitive Artificial Listener
SAM	Self-Assessment Manikin
SC	Skin Conductivity
SDK	Software Development Kit
SMO	Sequential Minimal Optimization
SNS	Sympathetic Nervous System
SVM	Support Vector Machine
SVR	Support Vector Machines for Regression
VLBP	Volume Local Binary Pattern

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction

Summary

In this chapter, the significance and future of affective computing are expressed. I focus on the computer vision techniques for recognizing human affective states and their potential applications. Current challenges facing facial expression recognition techniques are described and discussed briefly. The motivation of this thesis is also established in this chapter. I conclude this chapter by providing a walkthrough of the following chapters.

1.1 Introduction

Emotion and affective states play an important role in every-day life activities. As stated by Picard (2000), emotions control fundamental actions like decision-making, perception, rational thinking, and learning directly. In addition, some well-known scientists like Sigmund Freud even argued that emotions govern human behaviour (Werbos, 1994). More recently, psychologist and Nobel Prize winner Daniel Kahneman (2011) stated that emotion and mood can control the rational thinking.

Given the wide reaching impact of emotions and the fact that computers have become part of all aspects in our lives, engineers are looking into how the introduction of new types of electronic devices, such as smartphones and tablets, and new sensors like the Microsoft Kinect sensor provides more natural (human-like) and simpler ways of interaction with computers. In addition, the tendency of using digital devices is increasing over the world. According to International Telecommunications Union statistics (Sanou, 2013), the internet penetration among the world was 38% in 2013, and this rate is much higher in developed countries. Based on their estimation, 78% of the population in the developed world were using the internet in 2013. These numbers show the increase of new types of interactions between human and computers in everyday lives.

To enrich the Human-Computer Interactions (HCI), computers would be more useful if they could use these new types of devices to recognize the affective states first and then react accordingly. Emotional intelligence can maximise the naturalistic communication between human and computers to make a true user-friendly environment.

There are many applications for affect-aware interfaces. Affects and mental states, such as motivation, interest, and attention, are recognised as essential in learning with or without computers. Multiple authors have shown the impact of affect-aware systems in increasing the quality of learning (Calvo & D'Mello, 2011; S. Craig, Graesser, Sullins, & Gholson, 2004; R W Picard et al., 2004). Automatic affect detection is also important in other research areas, such as behavioural science, psychology, neurology, and psychiatry (R.W. Picard, Wexelblat, Clifford, & Clifford, 2002). For example, Affective Computing (AC) research has been used in the treatment process of people with autism spectrum disorders (Alzoubi, Hussain, & Calvo, 2014; C. Liu, Conn, Sarkar, & Stone, 2008).

Affective computing research tries to develop systems with two main abilities: perceiving and expressing emotions. Several methods and techniques have been proposed for sensing, detecting, and recognizing a human's affective states (Calvo & D'Mello, 2010; Zhihong Zeng, Pantic, Roisman, & Huang, 2009). Facial expressions, voice, posture, physiology, and text are the most investigated modalities for affect detection. Among the mentioned modalities, facial expression is the most common way to express emotions and regulate social interactions (Ekman & Rosenberg 2005). Even though a number of studies have been performed in the field of automatic Facial Expression Recognition (FER) (I. Cohen, Sebe, Garg, Chen, & Huang, 2003; Fasel & Luettin, 2003; Pantic & Rothkrantz, 2000b; Tian, Kanade, & Cohn, 2001), there is an interest to develop systems to detect more complex affective states (Baltrusaitis et al., 2011; Bixler & D'Mello, 2013; D'Mello & Graesser, 2010; Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013).

Physiology is one of the prominent modalities that has been used for affect detection because of its suitability for reflecting inner feeling and robust against deceptive behaviour. It has also been used in multimodal affect detection approaches (Hussain, Calvo, & Pour, 2011; Soleymani, Pantic, & Pun, 2012). Normally, physiological sensors need to be attached to the human body but this might be intrusive and make the application hard to adopt. Wearable sensors and devices were proposed to reduce the difficulties of setting up the traditional sensors. Among the current methods of measuring physiological signals, contact-less and remote methods are the most desirable. These methods are easy to adopt and cheaper than traditional devices (Poh, McDuff, & Picard, 2010). In addition, a remote contactless sensor could monitor several subjects at the same time.

Another approach for detecting affect is to combine multiple modalities (e.g., video and sound) that can be utilised in *Feature* or *Decision* levels. In the *Feature level fusion*, the extracted features from each modality are combined to make one big dataset. Then the classification task will be performed on this dataset. On the other hand, in the *Decision level fusion*, individual classifiers are applied on each modality and the final decision is made by analysing the classifiers' outputs. Both these approaches were used successfully in the affect detection application (Hussain & Calvo, 2011; Kim, 2007; Soleymani, Pantic, et al., 2012).

The goal of this thesis is developing a ***video-based*** method for detecting complex affective states in the naturalistic HCI. Cameras are used commonly everywhere, embedded in every electronic device. A video-based affect detection system can be adopted easily in many applications and environments. Our proposed method is focused on two main modalities that have been used for affect detection: *face* (from video) and *physiology*.

The feasibility of using video-based remote physiology sensors for affect detection is the first issue that needs to be evaluated. Although the remote sensing methods showed their performance for measuring physiological signals in controlled circumstances, the accuracy and robustness of these methods needs to be evaluated in more realistic scenarios.

The second issue relates to exploring the possibility of improving the affect detection by combining extracted physiological signals and facial features remotely. Dynamic and static facial features are extracted and evaluated for detecting affective states during controlled and naturalistic scenarios.

The third issue is the possibility of classifying ***complex affective states*** using extracted features from ***naturalistic*** HCI. Detecting complex affective states in the naturalistic scenario is a research area rarely explored (D'Mello & Calvo, 2013; M. E. Hoque, McDuff, & Picard, 2012). Most previous research tried to recognize basic emotions less likely to occur during an interaction between human and computers. Complex mental states like ***engagement***, boredom, curiosity, and frustration are the most reported states during HCI (D'Mello & Calvo, 2013). On the other hand, most of the previous attempts to build a prediction models relied on the posed affective states recorded in a controlled environment. Using these models to predict naturalistic affective states will not produce reasonable accuracy. To increase the generalizability of the system for detecting spontaneous affective states, natural affective states also need to be considered to build the prediction model (Afzal & Robinson, 2009).

1.2 Vision-based affect detection

The motivation of choosing video as the main modality for affect detection is the easiness of using video recording sensors at a low cost. A video sensor can be used for detecting and measuring facial expression, gesture, posture, and some physiological signals like heartbeats and respiration rates. It is expected that a reasonable accuracy can be achieved by considering all or a combination of these modalities using an ordinary camera. In this thesis, facial expression and heart rate (HR) features are considered as the two main inputs for training the models for affect detection.

As suggested in the literature, "*human face-to-face interaction is an ideal model for designing human-computer interaction systems*" (Pantic & Rothkrantz 2000). As indicated by numerous authors (Cowie et al., 2001; Pantic & Rothkrantz, 2003), humans mostly communicate emotions through facial expression compared with body gestures and vocal intonations. For example, a study by Mehrabian (1968) showed that the contribution of FER for affect detection was 55%, while they only relied on vocal utterances and spoken words by 38% and 7%, respectively. Another advantage of the facial expression is that most of the expressions are associated universally with specific affective states, particularly with basic emotions (Ekman & Friesen, 1978).

Previous studies showed that using just one modality for affect detection is not reliable (Jaimes & Sebe, 2007). Even humans do not rely on just one modality during a face-to-face interaction. Combining perceived information from different modalities and channels can make the HCI systems closer to the natural (human-like) interactions (D'Mello & Graesser, 2010; Pantic, Sebe, Cohn, & Huang, 2005). Each modality can cover the weakness of others for recognizing affective states.

Physiology is one of the most common modalities used in combination with other for affect detection (Hussain et al., 2011; Soleymani, Pantic, et al., 2012). Although it might less likely to be used in human-human interaction, there is potential for using physiological signals in affect detection. Physiological signals can reveal the internal feeling of the subject without the common issue of social masking or the common tendency to mask, hide, or pretend an emotional state due to social reasons. It is not an easy job to control and fake physiological responses to emotional stimulus (Kim & Andre, 2006; Peter, Ebert, & Beikirch, 2009). One of the main challenges with the physiological sensor is the need for physical contact with the subject but utilising contact-less sensors could address this.

The current challenges of video-based affect detection systems are related to compatibility with the real-world conditions and practical applications. Most of the systems reported so far have focused on ***posed*** and ***basic emotions***, which are not used commonly in practical AC applications. Another issue pertains to the ***real-time*** processing and affect detection, which is rarely addressed. Some of the challenges are also introduced because of the complexity of eliciting natural affective states (head movements, lighting conditions, automatic video segmentation, etc.). Despite the current notable progress in developing more naturalistic datasets in new application domains, more analysis and evaluation need to be performed to improve the accuracy and robustness of the systems. In the following sub-sections, the current challenges of vision-based affect detection and motivations of this thesis are discussed.

1.2.1 Video-based sensing of physiology

It might seem a bit strange, but research suggest that humans naturally detect physiological signals and can utilize them to recognize emotions in human-human interactions (RW Picard, Vyzas, & Healey, 2001). In most cases, close contact needs to be occur to sense, for example, clamminess during shake handing or heart pounding when someone sitting next to you. Sometimes it can be recognized remotely, for example, audiences can detect changes in the respiration rates of a lecturer (RW Picard et al., 2001). However, computers can provide more precise information about the internal feelings of humans, which can be useful for affect detection. The precise measuring of physiological signals could be considered as a significant advantage of the HCI over the human-human interaction for communicating emotions. Complex affective states, such as anxiety, depression, boredom, and frustration, can be recognized through analysing physiological patterns (Alzoubi, D'Mello, & Calvo, 2012; Healey & Picard, 2005; RW Picard et al., 2001).

On the other hand, the intrusiveness of traditional physiological sensors is always one of the main challenges for using physiology-based in real-world applications (Calvo & D'Mello, 2010). Another challenge relates to the large amount of the noise in the recorded signals due to movements of the users in the practical applications. Wearable sensors (Mann, 1997; Pantelopoulos & Bourbakis, 2010) and non-contact methods (Fei & Pavlidis, 2010; Li, Cummings, & Lam, 2009) for measuring physiological signals can be used to address the first challenge.

In the HCI application, using a vision-based method for monitoring physiological signals is a cheap and convenient method. The accuracy of these methods can also be improved using new machine learning techniques. However, these methods have been proposed recently and based on our knowledge; this is the first attempt to use a non-contact measurement of HR for affect detection.

1.2.2 Fusion model for affect detection

Multimodal affect detection techniques are becoming increasingly popular due to their better reliability and performance in detecting non-basic affective states (D'Mello & Graesser, 2010; Hussain & Calvo, 2011; Pantic & Rothkrantz, 2003; Soleymani, Pantic, et al., 2012). Naturally, humans use several modalities when they are interacting with each other. Each modality (face, voice, gesture, physiology, etc.) can represent unique aspects of each affective state. For example, previous studies showed that boredom and engagement as non-basic emotions cannot be detected easily using facial expression compared with other modalities (S. D. Craig, D'Mello, Witherspoon, & Graesser, 2008; McDaniel et al., 2007). Obviously, considering more modalities can increase reliability and accuracy of affect interpretation.

On the other hand, the way of representing non-basic emotions is different from person to person. This difference is even more challenging in the physiological patterns of emotions. Accordingly, building a reliable general model for affect detection using multichannel physiology still remains a challenging issue (Alzoubi et al., 2012). Adding more information from audio-visual modalities can improve the accuracy of affect detection.

Despite good progress in the field of multimodal affect detection, a few attempts have been made to combine physiology and other visual modalities like facial expression. Due to difference in the recording rate of these two modalities, it is hard to implement a fusion model in data or feature level. Finding a practical approach to combine different channels to achieve the best accuracy is still an open question for detecting naturalistic affective states (D'Mello & Kory, 2012). The ideal fusion model should provide a super additive performance, which has been rarely reported in the AC research. In this research, we explore the fusion of physiology channel (e.g., HR) and the facial expression channels in the feature level for recognizing non-basic affective states.

1.2.3 Naturalistic and none-basic affective states

As has been acknowledged in recent surveys (Calvo & D'Mello, 2010; Zhihong Zeng et al., 2009), there is a lack of practical datasets for affective computing that contain natural and non-basic emotions (M. E. Hoque et al., 2012). In addition, an ideal dataset needs to contain those affective states mostly common in AC applications. One reason that leads to this insufficiency is the complexity eliciting natural emotions. In most cases, a long session needs to be recorded and then analysed to obtain a set of valid emotional instances. Manual segmentation of natural affective states is a challenging and time-consuming task that also needs some prior knowledge. Concurrent experience sampling methods (Killingsworth & Gilbert 2010) might be a good solution to reduce the difficulties of gathering natural affective states. This thesis proposes a new method of concurrent experience sampling during HCI and evaluates the reliability of recorded samples using this method.

As the target of our proposed methods, we selected *Engagement* as a non-basic mental state. We selected *Engagement* because it is related strongly to the HCI applications (Peters, Castellano, & de Freitas, 2009) and it has been less targeted by automatic affect detection systems (D’Mello & Calvo, 2013). The authors of previous experiments have acknowledged the significance of detecting and maintaining engagement in learning and psychotherapy applications (Christenson, Reschly, & Wylie, 2012a; Kahn, 1990a). The ability of specific facial and physiological features for detecting user engagement during HCI needs to be evaluated (Hussain, Monkaresi, & Calvo, 2012b).

There is still a challenging debate about the methods of labelling and representing affective states (Russell, 2003). Categorical and dimensional representations are two common ways to quantify affect. Both models have advantages and disadvantages and finding a best representation of affect in practical AC applications needs more studies. However, dimensional models are considered to be more reliable (cross-lingual) for representing non-basic affective states that might not have equivalent translation in some cultures or languages.

1.3 Goals and approach of the thesis

The scope of this research is limited to vision-based automatic non-basic and naturalistic affect detection systems. The main objectives of this thesis are categorized in three main issues relate to vision-based affect detection techniques. Firstly, the feasibility of using video-based HR measuring in the HCI is evaluated. The limitation of current methods is evaluated and if it is necessary, a computational method is proposed to improve the accuracy of HR detection.

The second issue relates to the possibility of using remote HR sensing to improve non-basic affect detection in both controlled and naturalistic conditions. Several video-based methods for extracting facial features (geometric-based, appearance-based, and dynamic based methods) are implemented and utilized for affect detection. The aim is to explore the relation between each channel and each affective state and to find which channel or feature is more appropriate for non-basic affect detection. The performance of a multimodal approach by combining all extracted feature is also assessed in this thesis.

As the third issue, we aim to evaluate the possibility of training models for non-basic affect detection using naturalistic self-reporting.

1.4 Thesis contributions

The main contributions of this thesis are summarized:

1. Improvement of remote HR measurement in HCI: A key contribution of this thesis is evaluating the use of a contactless physiological method for affect detection that is less-intrusive and easy to adopt. Previous video-based contactless methods were not accurate for measuring HR during naturalistic HCI. A machine learning approach has been utilized to improve the accuracy of video-based HR detection in the HCI (Monkaresi, Calvo, & Yan, 2014).

2. Using remote sensing in affect detection: The proposed method for accurate remote sensing of HR is used to extract HR features for detecting non-basic affective states. The performance of these features in combination with other facial features were evaluated in two different studies (controlled (Monkaresi, Calvo, & Hussain, 2014) and naturalistic HCI).

3. Implementing geometric-based and appearance-based methods for affect detection: Geometric-based and appearance based features are the two main approaches used for recognizing facial expression. To evaluate the ability of the two common types of facial features for recognizing non-basic affective states, we have implemented two methods based on each approach. A geometric based method was implemented and tested on a previously recorded dataset (Monkaresi, Calvo, & Hussain, 2012). The second method, an appearance based method, was implemented and tested on the same dataset and a brief comparison reported (Monkaresi, Hussain, & Calvo, 2012a).

4. Developing and evaluating a fusion model of physiological signals and facial features: We developed a fusion model using multiple physiological channels and facial features (Monkaresi, Hussain, & Calvo, 2012b). The physiological signals were recorded using the sensors that needed physical contacts. The goal was to evaluate the feasibility of using the combination of multiple channels from different natures for detecting non-basic emotions. The results of this study encouraged us to explore the idea that using the contactless physiological sensors could also improve the accuracy of affect detection. In the latest study, the fusion model of remotely measured HR and two types of facial features (geometric-based and appearance-based features) was evaluated for detecting non-basic affective states in controlled and naturalistic conditions. A comprehensive comparison between the accuracies obtained by each of these channels for detecting arousal, valence, and engagement has been provided.

5. Building a new dataset for complex and naturalistic affective states during writing: Two protocols for affective data acquisition were proposed and utilized in this research. In the first one, a set of images from the International Affective Picture System (IAPS) (P. J. Lang, Bradley, & Cuthbert, 2008) has been used as emotional stimulus. The second one has been designed for a naturalistic HCI. A think-aloud method has been utilized for data labelling, and a set of interventions has been used during the recording sessions.

Other findings that are partially contributed by this research are:

1. Introducing a meta classifier for combining multiple classifiers for affect detection (Hussain et al., 2012b).

2. Comparing dimensional vs. categorical representations of affect in an automatic multimodal affect detection system during interacting with an intelligent tutoring system (Hussain, Monkaresi, & Calvo, 2012a).

1.5 Thesis overview

Chapter 2 starts with a brief description of different theories about emotions followed by a literature review on FER systems. A discussion is given on expressing the advantages of dimensional models of emotion over categorical models, particularly for modelling complex emotions. Some of the difficulties in collecting emotions in a naturalistic scenario, which is essential for validating emotion recognition systems, are also discussed in this chapter. A review on new computer vision techniques for extracting vital physiological signals from facial video recording is also presented.

Chapter 3 presents the methods and techniques used for recording, extracting features, and recognizing affective states. Three methods are introduced for feature extraction. The proposed framework for recognizing affective states by fusing three types of vision-based features is illustrated in the rest of this chapter. First, the Local Binary Pattern in Three Orthogonal Planes (LBPTOP) is introduced as a dynamic texture-based method to extract appearance-based and motion-based features. Second, the Microsoft Kinect face tracker engine (SDK v. 1.5) is used to extract geometric-based features. Third, a new method for recovering the HR using computer vision techniques is presented. Two main experiments designed and conducted to elicit facial expression in a naturalistic environment are described in this chapter.

Chapter 4 evaluates the accuracy of our proposed method for the remote measuring of HRs. The accuracy of the proposed method is reported in three different experiments and the achieved results were compared with the actually recorded HR using an electrocardiograph (ECG) device.

Chapter 5 reports the results for automatic affect detection using three channels described in Chapter 3 during a controlled HCI. A dimensional representation of affect in two dimensions (valence/arousal) is used in this study. A mapping between dimensional affect and categorical representation of affect is also provided in this chapter. User-dependent, gender-specific, and user independent analysis are also evaluated.

Chapter 6 presents the results of automatic affect detection in the third study, which is a naturalistic HCI. In this study, the goal is detecting engagement as a complex affective state during a writing session using a computer. Besides engagement, the accuracies of the proposed methods for detecting dimensional affect are also reported. The impact of feedbacks and interventions on the engagement level is explored.

Chapter 7 concludes by describing the principal outcomes and contributions of this thesis. The limitation of our proposed method and some suggestion for future work are also presented.

Chapter 2. Background and literature review

Summary

Psychological theories of emotion provide theoretical grounding for the development of affective computing technologies. This chapter provides a brief description of different theories about emotions used in AC, along with a literature review on FER systems. A discussion is given on expressing the advantages of dimensional models of emotion over categorical models. The importance of implementing systems for recognizing complex emotion over basic emotion is discussed. Current attempts in the field of affective computing are suffering from lack of enough reliable datasets for training and testing their systems. Data gathering (e.g., emotion elicitation) is still a challenging issue in this area. This issue is particularly challenging when gathering naturalistic emotions and some of the difficulties in collecting emotions in naturalistic scenarios are discussed in this chapter. We have introduced current FER works and evaluated some of the most popular systems. We also give a review on new computer vision techniques for extracting vital physiological signals from facial video recording.

2.1 Theories of emotions

Different definitions have been introduced for affect. In each discipline (e.g., psychology, social science, neuroscience) it has been defined with specific characteristics from different perspectives. Affect often has a broad and ambiguous definition, which includes other aspects, such as emotions, feelings, moods, and attitudes (Calvo & D'Mello, 2010). Emotion could be considered as the core of the affective phenomenon. The definition of mood also has common features of emotion with some distinctions.

Like emotion, mood is also a subjective experience, which is expressed using common human communicational channels (Moridis & Economides, 2009). The differences between emotion and mood can be described by parameters, such as duration, timing, and cause-reaction (Larsen, 2000). However, studying different theories about emotions can give us a better insight into this fuzzy and complex phenomenon. Current theories that try to explain emotion are divided to six categories and discussed in the current survey. Four of these theories are derived from traditional theories of emotion that consider emotions as expression (Darwin, 1872; Frijda, 1987), embodiment (James, 1884), cognitive appraisal (Arnold, 1960), or social construct (Averill, 1980). The other two have been introduced later and consider emotion as an outcome of neural circuitry (Dalgleish, Dunn, & Mobbs, 2009) or as a psychological construction that integrates different perspectives of emotion (Russell, 2003). In the rest of this chapter, we introduce two common theories of emotions relevant to this research.

2.1.1 Emotions as expressions

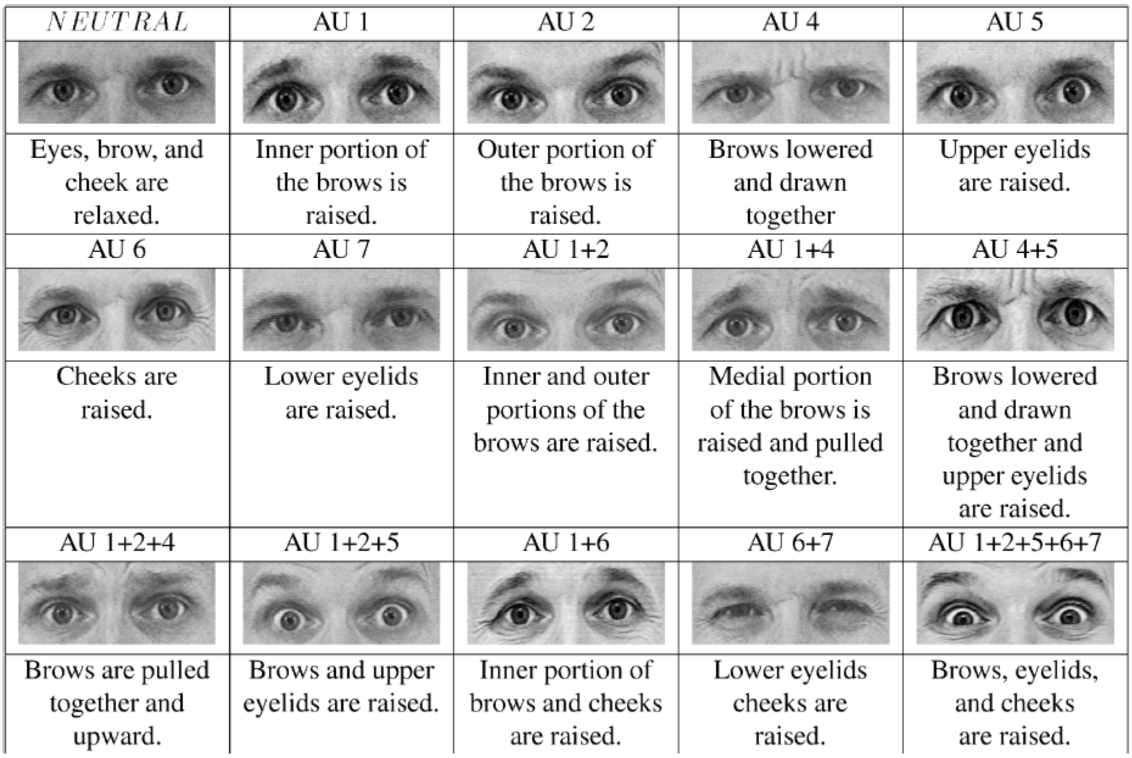
Expressions are the most common and generally accepted feature of emotions. Most emotions are communicated through expression, such as facial expression and body movement. Charles Darwin (1872) was the first scientist to explore the emotion from the expression perspective. Inspired by his evolutionary theory, he noticed that there are similarities between human and animals in terms of some of the basic facial and body expressions. Other researchers extended his theory and argued that an action is associated with each emotion expressions (Frijda, 1987). Accordingly, to define each emotion category, we need to find an explicit action (or tendency) associated with it. For example, to define "fear" as an emotion expression, the action of "avoidance" was proposed.

Although this definition of emotion could not explain some emotional behaviours, it has been used widely for implementing AC systems. Current technologies allow us to record and analyse facial and body expressions to recognize emotions. A large portion of research in affect detection has been devoted to analysing facial expressions.

2.1.1.1 *Facial Action Coding System*

Ekman and Friesen (1978) proposed the most common method for modelling facial expressions has and this has been called Facial Action Coding System (FACS). They introduced a set of 44 facial action units (AUs) that can represent different emotions. In fact, each AU can be identified by measuring the movement of specific facial component. For example, AU1 means the inner portion of the brows is raised and AU12 represents a lip corner pull. Figure 2-1 shows the descriptions of upper face AUs in FACS. As illustrated in Figure 2-1, the AUs can be considered in isolation or in combination with other AUs. According to Scherer and Ekman (1982), more than 7000 combinations of AUs have been observed. Each combination can also be in two forms: *additive* or *non-additive*. In the additive combination, the appearance of the AUs will not change after the combination. However, in the non-additive combination, the appearance of AUs will change due to combinations (AU1+2 and AU1+4 are two examples of non-additive combinations). So, recognising non-additive AUs is much more difficult than recognising additive combinations.

FACS only presented a coding schema for facial movements and does not provide a mapping between AUs and specific emotions. Due to its power in describing details of facial movements, this coding system has been used successfully in AC applications. Manual coding of facial AUs is a difficult and time consuming task, and expertise is needed to recognize combinations of AUs during expressing emotions. Training of a certified coder needs almost 100 hours. In practice, manual extracting of AUs from one minute of recorded video needs approximately one hour (Donato, Bartlett, Hager, Ekman, & Sejnowski, 1999). Another issue with this coding system is that the FACS was designed for static images and it is difficult to adopt it in a dynamic space.



NEUTRAL	AU 1	AU 2	AU 4	AU 5
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 2-1: Descriptions of upper face AUs in FACS from Tian et al. (2001)

To reduce the difficulties associated with the FACS, automatic AU detection systems have been proposed (Tian et al., 2001; Tong, Liao, & Ji, 2007; Valstar & Pantic, 2012; Whitehill &

Omlin, 2006). The first step in FACS-based affect detection systems is recognizing AUs. Afterwards, machine learning techniques (Neural networks, Hidden Markov Models, Bayesian networks, etc.) are utilised to predict observed affective states using extracted AUs.

Besides FACS-based methods, many other methods and techniques have been proposed for FER. A comprehensive review on two common approaches for FER systems (geometric-based and appearance-based methods) is discussed in Section 2-4. Naturally, most of the introduced systems for facial and body expression recognition have relied on visual modality. In this thesis, we also use the theory of *emotions as expressions* to extract facial features and then recognize emotions.

2.1.2 Emotions as embodiments

Considering emotions as embodiments of peripheral physiology is another theory for explaining emotion. According to this theory (James, 1884), physiological changes are the main reasons for experiencing emotions. For example, feeling “fear” is associated with the acceleration in heart rate. By identifying specific patterns of physiological changes, we can recognize each associated emotion. This theory also known as *James–Lange Theory* (Cannon, 1927) and is investigated under the topic of *psychophysiology*.

Damasio (2008) had a similar argument that assumes the Sympathetic Nervous System (SNS) and Parasympathetic Nervous System (PNS), the two parts of the Autonomic Nervous System (ANS), are the main source of feeling emotions. The sympathetic part is responsible for preparing the body for high levels of physical activities during emergency or stress situations. However, the parasympathetic part is trying to keep the body in the state of the rest and relaxation (Andreassi, 2007). Accordingly, in response to each emotional stimulus, the SNS and PNS produce certain physiological changes in the body. These physiological changes are interpreted as emotions in the brain. Based on this theory, for each emotion specific patterns of ANS, activities can be observed. Electrocardiogram (ECG), Electromyogram (EMG), Skin Conductivity (SC), and Respiration (RSP) are the main methods for measuring the ANS activities.

By introducing new technologies for monitoring different kinds of physiological signals, this theory has been received more attention of academic studies. Many AC systems have utilized physiological signals to recognize specific affective states. The main concern about these systems is the intrusiveness of sensors used for recording physiological signals (Calvo & D'Mello, 2010). However, this issue has been addressed by introducing wearable sensors (Arroyo et al., 2009) and remote sensing techniques (Poh, McDuff, & Picard, 2011).

2.2 Dimensional vs. categorical models of emotions

Affective states can be represented through either categorical or dimensional models and each representation has advantages and disadvantages. The categorical representation considers emotions as discrete and independent categories. This representation was inspired by the concept of basic emotion introduced by Darwin (1872) and then extended by Ekman and his colleagues (Ekman 1992; Ekman & Friesen 1971). According to Ekman and Friesen (2003), human can recognise six basic emotions universally and cross-culturally: anger, disgust, fear, happiness, sadness, and surprise. Other researchers also specified up to 18 categories as basic emotions (Ortony & Turner, 1990). However, Ekman's basic categories have been explored widely in AC research.

Despite the ability of categorical representation in modelling basic emotions, it did not show a good performance in representing complex affective states or blended emotions (Gunes & Pantic, 2010; Yu, Aoki, & Woodruff, 2004). Discrete categories of emotions cannot describe emotions that occurred mostly during every-day human interactions (Zhihong Zeng et al., 2009). Sometimes it is difficult to specify a rigid border between two complex affective or mental states. Another drawback of categorical representation of emotion is that it is not always possible to translate an emotion category from one culture (language) to another one. For example, there is no true equivalent for “disgust” in Polish (Russell, 1991).

To address these issues of the categorical representation of emotions, dimensional models have been proposed. In this representation, an affective state is described using one or multiple numeric dimensions. Two most common dimensions that reflect the main aspects of emotions are *valence* and *arousal*. Valence represents the level of pleasure of each affective state and ranges from highly negative (e.g., unpleasant) to highly positive (e.g., pleasant) feelings. Arousal indicated the level of activation of each affective state that ranges from passive (e.g., sleepiness or boredom) to active (e.g., frantic excitement). Researchers have considered other dimensions like potency (dominance), control, power, and expectation (Mehrabian & Russell, 1974). By using these dimensions, different models of affect can be described. For example, Russell (1980) developed a two dimensional (2D) model of affect using valence and arousal, as shown in Figure 2-2.

Both categorical and dimensional models of affective states have received substantial attention in the AC research area. Early attempts mostly focused on labelling emotional categories (Banse & Scherer, 1996; Kanade, Cohn, & Tian, 2000; Pantic, Valstar, Rademaker, & Maat, 2005), whilst new affect databases include dimensional annotations (Douglas-Cowie et al., 2007; Soleymani, Lichtenauer, Pun, & Pantic, 2012). Soleymani et al. (2012) created a multimodal dataset for affect detection and video implicit tagging. Beside emotional labels, they have elicited arousal, valence, dominance, and predictability during self-reporting. A Sensitive Artificial Listener (SAL) database (Douglas-Cowie et al., 2007) is a public audiovisual database that has been labelled using valence and arousal dimensions.

The audiovisual information was recorded during interaction with a SAL in a Wizard-of-Oz scenario.

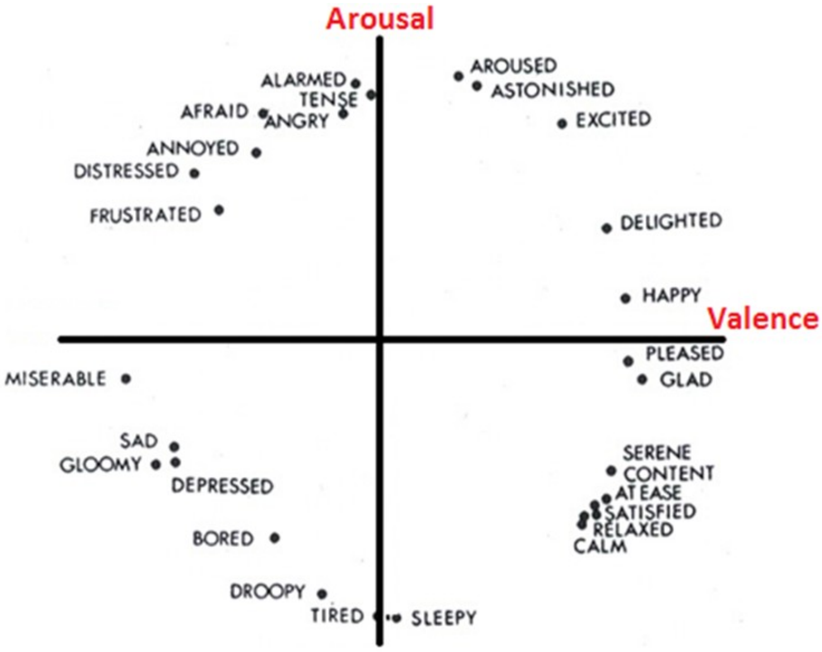


Figure 2-2 : A two dimensional representation of affective states. Adopted from **Russell (1980)**

2.3 Emotion elicitation

Affect elicitation is one of the most challenging issues in affect recognition. Conducting an experiment to create a dataset for affect detection is an expensive, time-consuming, and difficult task. Several techniques have been proposed for affect elicitation. Due to difficulties associated with collecting information in realistic environment, most of the affect elicitation experiments take place in laboratories. There is still an open discussion regarding the possibility of collecting real emotions in a laboratory (Afzal & Robinson, 2009; R Picard, 2003). However, designing an experiment protocol for HCI applications to induce genuine and spontaneous affective states is still a challenging issue. This issue is more difficult for inducing complex affective states.

Picard et al. (2001) addressed the main concerns that need to be considered during emotion elicitation. The stimulus used for inducing emotions is an important factor in emotion elicitation methods. Different stimulus, such as event, image (P. J. Lang et al., 2008), music (Lichtenstein, Oehme, Kupschick, & Jürgensohn, 2008) or movies (Soleymani, Pantic, et al., 2012), have been used to trigger emotions. Awareness of the subject about the purpose of the experiment might have an impact on the reliability of the elicited data.

According to a recent survey (M. E. Hoque et al., 2012), there is still no dataset that contains spontaneous non-basic emotions that is ideal for affect recognition systems. Figure 2-3 shows the position of current state of the art datasets in terms of spontaneity and complexity of emotions. Seven datasets are introduced in this figure: Spaghetti (Douglas-Cowie et al., 2007); SAL (Douglas-Cowie et al., 2007); Semaine (Mckeown, Valstar, Cowie, & Pantic, 2010); MMI (Pantic, Valstar, et al., 2005); RU-FACS (Bartlett et al., 2006); and Mind Reading (Rana El Kaliouby, 2005). Spaghetti, SAL, and Semaine provided spontaneous affective states but they focused mostly on basic emotions. On the other hand, RU-FACS and Mind Reading datasets contained more complex affective states while they were mostly acted.

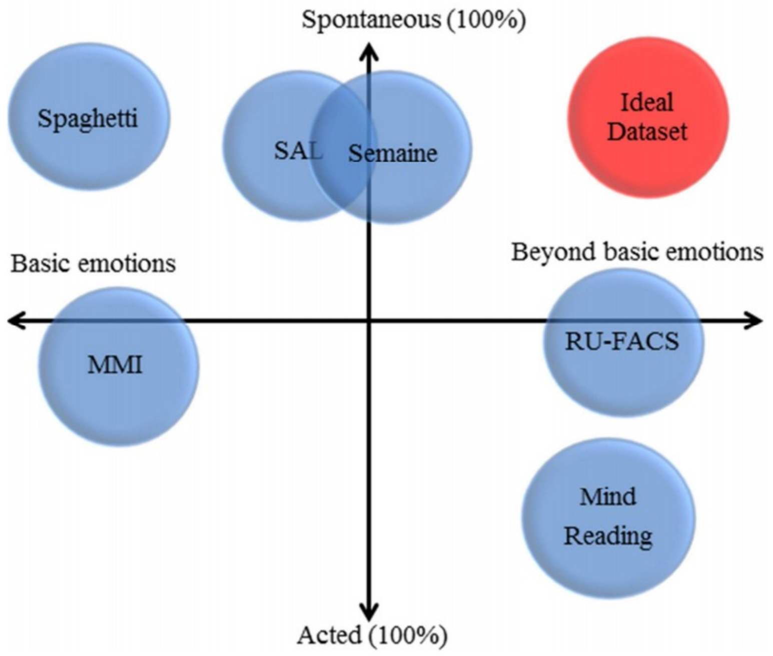


Figure 2-3 : Comparison of current state of the art dataset by considering two dimensions: (spontaneous vs. acted) (basic vs. non-basic) emotions. From Hoque et al. (2012)

Table 2-1 presents a list of datasets currently used for affect recognition systems. The MMI dataset (Pantic, Valstar, et al., 2005) is one of the public datasets used commonly for evaluating different video-based affect detection systems. Compared with other datasets, this dataset contains a large number of video recordings of subjects from different cultures. Most of the videos were recorded in both frontal and profile views of the face. Although it contains few samples of spontaneous facial expressions, more than 85% of the facial expressions were acted deliberately. Another issue associated with this dataset is lack of complex affective states. The Mind Reading dataset (Rana El Kaliouby, 2005) was an early attempt to create a visual dataset of complex mental states. Several non-basic mental states, such as thinking, concentrating, confused, interested, unsure, agreement, and disagreement, exist in this corpus. However, actors posed all of these facial expressions.

On the other hand, Semaine, SAL, and Cam3D datasets tried to record more naturalistic and spontaneous affective states. These three datasets were recorded during dyadic conversations.

The SAL dataset consists of audio-visual recordings of a pretended human-computer interaction (Wizard-of-Oz approach). The SAL interface contains four characters with four different personalities. Interacting with each of them can induce specific emotions of the participants: happy, sad, angry, and pragmatic. The data were also annotated using a 2D scale (valence/arousal).

In the Semaine dataset, subjects interacted with an operator who tried to evoke emotional reactions from them. The recorded conversations were annotated for five affective dimensions: valence, activation, power, anticipation/expectation, and overall emotional intensity. In addition to the dimensional annotations, categorical labels were also provided for each instance.

The Cam3D provides a three dimensional (3D) dataset of facial expressions, including the upper body. This corpus contains 108 audio-visual segments of natural complex mental states recorded during dyadic conversations. The 3D models of the face and the upper body were created using the Microsoft Kinect sensor. The recorded videos were segmented manually and annotated using crowd-sourcing method. They used categorical labels rather than dimensional ones because they asked naive labellers for crowd-sourcing (Mahmoud, Baltrusaitis, Robinson, & Riek, 2011).

Table 2.1: List of available datasets for affect recognition

	Elicitation method **S: Spontaneous** **P: Posed**	**Modalities** **V:Video** **A: Audio**	**Size** **(subject/instances)**	**Emotion description** **C: Categories** **D: Dimensions**
Cam3D (Mahmoud et al., 2011)	Spontaneity: S	A/V	Subjects: 7 Instances: 108	Basic and Complex Labelling: **C** (12 categories)
CK+ (P. Lucey et al., 2010)	Spontaneity: P and S	V	Subjects: 210 Instances: 700	Basic emotions and AUs Labelling: **C** (6 categories)
Semaine (Mckeown et al., 2010)	Spontaneity: S	A/V	Subjects: 20 Instances: 578 duration 6:30:41	Basic emotions Labelling: **D** (5 dimensions) and C (6 categories)

SAL (Douglas-Cowie et al., 2007)	Spontaneity: S	A/V	Subjects: 24 duration: 4:11:00	Labelling: **C** and D
RU-FACS (Bartlett et al., 2006)	Spontaneity: P	A/V	Subjects: 100 Instances: N/A	AUs Labelling: **C**
MMI (Pantic, Valstar, et al., 2005)	Spontaneity: P and S	V	Subjects: 79 Instances: 2894	Basic emotions and AUs Labelling: **C** (6 categories)
Mind Reading (Rana El Kaliouby, 2005)	Spontaneity: P	V	Subjects: 30 Instances: 1742	Complex mental states Labelling: C
GMEP (Bänziger & Scherer, 2010)	Posed (professional actors)	A/V	Subjects: 10 Instances: 7000 (1260 were annotated)	AUs, Labelling: C (18 emotion)

2.3.1 Segmentation

The observation time for annotation is an important parameter that needs to be specified based on the context of the recorded dataset. The annotation could be performed either continuously or discretely. Two main approaches for segmentation are *interval-based* and *event-based*. Different factors need to be considered for choosing the best approach for segmentation. The subtleness of reported emotions, complexity of annotation scheme, and the frequency of expression occurrence are the most important factors to choose between *interval-based* and *event-based* approaches (Afzal & Robinson, 2009).

In interval-based methods, the coder needs to report in certain time intervals. Each segment might include one or more affective states, which can increase the complexity of annotation task. The size of intervals also depends on the context of the system and annotation scheme. For example, the annotation scheme should allow reporting multiple emotions in more details. The *interval-based* segmentation can be done automatically but for event-based segmentation, we need an intelligent system expert to recognize appropriate events. Each event might contain a facial expression, head, or body movement or a combination of these behaviours. Although manual event-based segmentation is a complex and time-consuming task, it can increase the quality of annotations (Mahmoud et al., 2011).

2.3.2 Annotators

One important aspect to create a reliable dataset for affect detection is selecting the appropriate annotators. Annotation task can be performed by the subject (e.g., learner or user) or by an external annotator (e.g., expert, peer, tutor, etc.). Both types of annotations have advantages and disadvantages and there is still a debate between pros and cons of each of those methods to create a reliable dataset (Porayska-pomsta, Mavrikis, D'Mello, Conati, & Baker, 2013).

Self-reports: Self-reporting is a common method for assigning emotional labels to each recorded instance. This method can represent the internal feeling of the subject perfectly; however, the quality of the self-reports depends upon two critical factors: the ability of the subject to report his/her feeling, and the complexity of emotions. Age, cultural background, and personality of the subject influence the ability of reporting internal feeling. For example, researchers showed that children under age of 8 had a naive understanding about emotions, particularly complex emotions (Conati & Maclaren, 2009). Accordingly, these factors need to be considering when the annotation scheme (e.g., questionnaires, emotion models, etc.) is designed. The most important issue associated with this method is creating subjective data. Each person has a specific personality with specific interpretation of emotion and internal feelings. Training and building general models using subjective self-reports is always a challenging task.

External annotators: External annotators can be a peer, tutor, expert or non-expert judges. Similar to self-reports, age and cultural proximity (Elfenbein & Ambady, 2003) to the subjects are the most important factors to be considered for selecting annotators. Familiarity with the context of the recorded data is also another important factor that can impact the quality of the produced labels. For example, de Vicente (2003) showed that postgraduate annotators could annotate their classmate interactions more accurately as indicated by high inter-rater agreements. On the other hand, younger and less experienced annotators (e.g., undergraduate students) might produce annotations with poor inter-rater agreements. Most researchers agreed that training the annotators for assessing emotions is essential (Sayette, Cohn, Wertz, Perrott, & Parrott, 2001).

The main advantage of using external annotators is avoiding subjective differences and creating smooth annotations, because all instances will be assessed by the same annotators with the same constructs (Porayska-pomsta et al., 2013). However, compared with the self-reports, the original internal feeling is less reflected through the external annotations.

2.3.3 Concurrent and retrospective annotation

The annotation tasks (self-report and external annotation) can be conducted concurrently or retrospectively. In the concurrent annotation methods, the reports are collected when the affective states are recorded. Different instruments and techniques have been used for concurrent annotation based on either *free-response* techniques, such as *think-aloud* and *emote-aloud* protocols (S. D. Craig et al., 2008), or *forced-response* techniques. Though these techniques can be adopted for both types of annotators, using concurrent methods for external annotators is less common (Porayska-pomsta et al., 2013). Collecting self-reports concurrently can increase the quality of annotations because in-the-moment insights of the subjects can be captured. On the other hand, the risk of cognitive load implications is associated with the concurrent annotation methods.

The retrospective annotation is much easier to prepare, adopt, and administer compared with concurrent annotation. In this method, the annotation reports are collected after the recording sessions through interviews or forced-choice questionnaires. The annotators can watch the recorded video/audio as much as they want and focus on assessing the observed emotion in more details. This approach is also appropriate for external annotators. As another advantage, this method does not produce any additional cognitive load during the recording sessions. Accordingly, creating a dataset using this method needs extra time for annotation compared with the concurrent method.

2.4 FER systems using image and video classification

Similar to other pattern recognition systems, FER systems consist of two major components. The first and the most important component for FER systems is the facial feature extraction component (Fasel & Luettin, 2003). According to the literature (Koelstra, Pantic, & Patras, 2010; Zhihong Zeng et al., 2009), current approaches for feature extraction are divided into two main categories: *Geometric-based* and *Appearance-based* approaches and these are described in more details in following sub-sections. The second component is the classification component that utilizes various types of classifiers or probabilistic models to predict the affective states or AUs.

2.4.1 Geometric-based approach

Geometric-based methods rely on extracting some features by tracking a set of fixed points or landmarks on the face for affect recognition. In the early attempts, these facial landmarks were placed manually on the face and then features were extracted by measuring the changes in the position of each point compared with the neutral face. However, current systems use computer vision techniques to detect and track face and other facial components automatically. A list of FER systems that used geometric-based methods for feature extraction is presented in Table 2-2.

The first group of the geometric-based methods tracks the position of some fixed points on the face (Baltrusaitis et al., 2011; R. El Kaliouby & Robinson, 2005; Pantic & Patras, 2006; Valstar & Pantic, 2007). Baltrusaitis et al. (2011) proposed a new real-time method for affect recognition based on a system implemented by El Kaliouby and Robinson (2005). Besides 22 facial points, they tried to classify head and body gestures using Hidden Markov Model (HMM). In the next step, they used a multi-level dynamic Bayesian Network to identify each affective state. They trained and test their system on a new dataset GEMEP-FERA (Valstar, Jiang, Mehu, Pantic, & Scherer, 2011) and achieved an average classification accuracy of 44% that was less than its former implementation. El Kaliouby and Robinson (2005) achieved the classification accuracy of 77% when they validated their system using a dataset produced by themselves.

Table 2.2: Geometric-based systems for affect/AU detection

	Ref	Approach/ Method	Classifier	Dataset Participants	Emotion model/ labels	Classification Accuracy
1	El Kaliouby and Robinson (2005)	G:1 Facial points 24 facial points	HMM, DBN	Own dataset 30 subject	6 Complex mental states	77%
2	Baltrusaitis et al. (2011)	G:1 Facial points 22 facial points + gestures	HMM, DBN	GEMEP-FERA	6 Emotion categories	44%
3	Pantic and Patras (2006)	G:1 Facial points	Rule based	MMI: 19 subjects	27 AUs	86.3%
4	Valstar and Pantic (2007)	G:1 Facial points 20 facial points	GentleSVM	MMI: 52 subject	2 Classes Posed vs. spontaneous smiles	94%
5	Chang et al. (2006)	G: 2 Shape model 58 facial landmarks	Probabilistic method	Own dataset 2 subjects	6 basic emotions	–
6	Kotsia and Pitas (2007)	G: 2 Shape model	SVM	CK	6 basic emotions	99.7%
7	Tian et al. (2001)	G:3 Facial components Lips, eyes, brows, cheeks and furrows	Neural Network	CK	16 AUs	Upper face: 96.4% Lower face: 96.7%
8	Valstar and Pantic (2012)	G:3 Facial component 20 facial points	GentleSVM, HMM	CK, MMI	22 AUs	Posed: 95.3% Spontaneous: 72%
9	Gunes and Pantic (2010)	G: Head gestures	SVR	SAL	5 dimensions	–

The second group of the geometric-based methods needs to build a facial shape model to extract geometrical features (Chang et al., 2006; Kotsia & Pitas, 2007). Chang et al. (2006) proposed a method to map facial expression into a low-dimensional space. They created an active shape model for tracking 58 facial points and then converted them into a 3D space using the Lipschitz embedding method. They used a dataset with only two subjects that was not sufficient to create a subject-independent model.

Kotsia and Pitas (2007) proposed and evaluated two methods using a facial grid model. The grid needed to be assigned manually to some depicted facial landmarks at the first frame. In their first method, they tracked the facial grid and used a multiclass SVM to predict the emotion. In their second method, they tried to classify deformations of facial grid using

SVMs to identify specific AUs and then recognize the six basic emotions using certain rules applied on detected AUs. They achieved slightly better accuracy when they used the first method (99.7% vs. 95.1%).

The third group of the geometric-based methods tries to model facial components and extract geometric features from that model (Tian et al., 2001; Valstar & Pantic, 2012). Eye, lips, brows, and furrows are the most common components used for affect detection. As shown in Figure 2-4, various geometric features can be measured using the three facial components: eyes, brows (white lines above eyes), and cheeks (black solid lines under eyes). For example, the height of the left eye can be measured by this equation: *(hl1+hl2)*. The distance between two eyebrows is another feature that can be measured by D (see Figure 2-4). These features can be considered as the inputs of a classifier to identify specific AU or a facial expression. Tian et al. (2001) used a neural network-based recognizer to identify 16 AUs using these geometrical features.

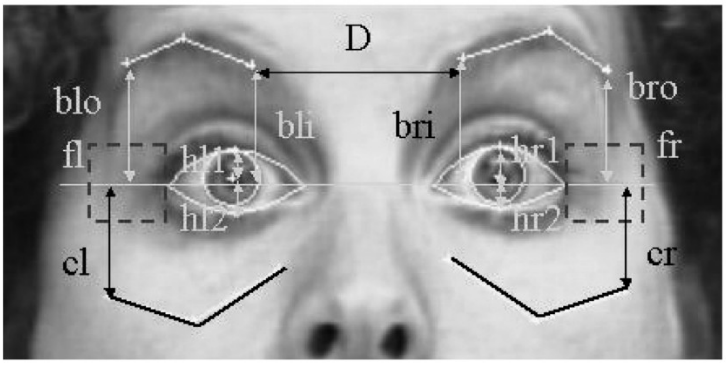


Figure 2-4: An example of upper face geometric-based features. From: Tian et al. (2001)

As shown in Table 2-2, most of the geometric-based systems have used categorical representation of emotions. Dimensional emotion prediction using geometric-based features has been rarely explored in the FER research area. Gunes and Pantic (2010a) proposed a method to predict emotion in five dimensions (arousal, expectation, intensity, power, and valence) from head postures. They applied their system on a dataset that contains spontaneous facial expressions and head gestures. Their results demonstrated that building a user-independent model for detecting dimensional emotions is feasible.

2.4.2 Appearance-based approach

In the appearance-based approach, the motion and deformation of certain regions of the face (e.g., skin or texture) are considered to extract facial features. Some facial actions hardly detected by geometric method, such as wrinkles, bulges, and furrows, can be recognized easily using appearance-based techniques. Researchers used different appearance-based techniques, such as Gabor wavelets (Guo & Dyer, 2005; Littlewort et al., 2011; Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2006); Optical flows (Anderson & McOwan, 2006); Local Binary Patterns (Moore & Bowden, 2011; Shan, Gong, & McOwan, 2009); Active Appearance Models (S. Lucey, Ashraf, & Cohn, 2007); and Negative matrix factorization (Zhi, Flierl, Ruan, & Kleijn, 2011) for recognizing AUs and affective states. Most of these attempts have been focused on recognizing either AUs or basic emotions. Table 2-3 briefly summarizes the FER systems that attempted to use appearance-based methods techniques together with the best results reported.

Table 2.3: List of Appearance-based systems for affect/AU detection

	Reference	**Approach/ Method**	**Classifier**	**Dataset Participants**	**Emotion model**	**Result (Accuracy)**
1	Guo and Dyer (2005)	A: Gabor	Bayes Classifier, AdaBoost, SVM	Other: 10 Japanese women	Seven basic emotions	SVM:92.4%
2	Bartlett et al. (2006) Bartlett et al. (2005)	A: Gabor	AdaBoost SVM	CK+EH:119 RU:12	AUs and Basic emotions	Acc: CK:93.4 RU: 90.5
3	Anderson and McOwan (2006)	A: Optical	SVM	Other: 100 students	Six basic emotions	75.32%
4	Littlewort et al. (2011)	A: Gabor	SVM	CK+	19 AUs Six basic emotions	AUs: 90.1% Emotions: 80%
5	Shan et al. (2009)	A: LBP	SVM	CK	Six Basic emotions	91.5%
6	Zhi et al. (2011)	A: NMF	–	CK	Six Basic emotions	CK: 94.3%
7	Lucey et al. (2007)	A: AAM	SVM	CK: 100 subjects	AUs	Acc: 95%
8	Asthana et al. (2009)	A: AAM	SVM	CK	Seven basic emotions	Acc: 94.3%
9	Zhao and Pietikäinen (2007); Zhao and Pietikäinen (2009)	A: DT, LBP	SVM	CK	Six basic emotions	10 fold: 96.26%
10	Koelstra et al. (2010)	A: DT	GentleBoost, HMM	MMI CK	AUs	MMI (27 AUs): 94.3% CK (18 AUs): 89.8%
11	Tian et al. (2002)	A & G	Neural Networks	CK	9 AUs	Geometric: 87.6% Combine: 92.7%
12	Shin (2007)	A: locally linear embedding (LLE)	NN	Other: Korean dataset	2 dimensions, 4 classes	–
13	Grafsgaard et al. (2013)	A: using CERT toolbox (Littlewort et al., 2011)		Other	Engagement, Frustration	Regression models were proposed

The Gabor-wavelet-based method is one of the early appearance-based methods for facial expression recognition. Zhang et al. (1998) shown that the Gabor-based method outperforms the geometric-based method for facial expression recognition. Guo and Dyer (2005) selected 34 facial points manually after applying Gabor filters. The calculated amplitudes of each fiducial point were used for classifying seven basic emotions. Barlett et al. (2006) first detected the face and eyes and then used Gabor wavelet filters to extract appearance-based features. They extracted 165,888 features using Gabor filters in eight orientations and nine frequencies. For classification, they evaluated two classifiers (SVM and Adaboost) and figured out that the best results can be achieved when the AdaBoost was used for feature selection and the SVM classifier was used for training.

Using optical flow of the face to extract facial features is another common method for affect detection. Optical flow is a method to represent motions of facial components. Different patterns in facial motions can determine specific affective states. Anderson and McOwan (2006) used the multichannel gradient model to extract optical flow features by measuring the velocity of different facial regions. They used their system to detect six basic emotions and evaluated it in an emotional chat-room and a desktop application.

Negative matrix factorization (NMF) is a new appearance-based technique that has been used recently for emotion detection. This method tries to calculate a sparse representation of the face. Zhi et al. (2011) showed that it can recognize successfully six basic emotions even if some parts of the face were occluded. Facial occlusion is one of the big issues for automatic FER that frequently occurs in naturalistic scenarios. Zhi et al. (2011) selected a set of facial images from the CK dataset that were partially occluded and evaluated the performance of their system using those images. The detection accuracies of 93.3%, 94%, and 91.4% were achieved for eyes, mouth, and nose occlusions, respectively.

Dealing with individual differences needs to be considered during designing a affect detection system. On the other hand, an ideal affect detection system should be able to generalize the prediction model to previously unseen persons. Chu et al. (2013) proposed a solution to personalize a generic classifier using an unsupervised method. They have utilized a Selective Transfer Machine (STM) to reweight more the training instances that are closer to the test instances. The STM is a classifier-independent technique and can be used with any classifier. They have used the STM with SVM classifier and the results showed that the STM technique outperformed generic classifier for detecting eight most frequent action units (Chu et al., 2013).

As reflected in Tables 2-2 and 2-3, most previous works focused on detecting basic emotions, but more recently, some researchers have focused on the recognition of complex mental states, such as engagement (Bohus & Horvitz, 2009; Grafsgaard et al., 2013), attention (Roda & Thomas, 2006; Vertegaal, 2003), and common affective states in learning applications (D'Mello & Graesser, 2010; Graesser et al., 2006; McDaniel et al., 2007). Instead of relying on single modality, most of the these researchers try to recognize more complex affective states using multiple cues, such as facial expression, head, and body posture and audio, text, and physiological signals.

Grafsgaard et al. (2013) used the computational toolbox proposed by Littlewort et al. (2011) to track facial movements within a naturalistic video corpus of naturalistic tutorial dialogue. The most frequent facial AUs, including eyebrow raising (inner and outer), brow lowering, eyelid tightening, and mouth dimpling, were selected to predict engagement, frustration, and learning gains using forward stepwise linear regression. Engagement was measured through three different aspects of engagement, such as mental demand, physical demand, temporal demand, and performance, using a post-session survey. Their findings suggested that upper face movements would be a good predictor of engagement, frustration, and learning, and they achieved reasonable agreement between their predictions and manual annotations (Grafsgaard et al., 2013).

2.4.2.1 *Dynamic texture-based techniques*

One of the new methods introduced in the appearance-based approach is the Dynamic Texture (DT)-based method (Chetverikov & Renaud, 2005; Saisan, Doretto, & Wu, 2001).

DT is a specific pattern that can be observed in a spatiotemporal space. Typical examples of DTs are smoke, fire, and sea waves. Facial expressions can be considered as DTs. Each expression has certain temporal facial features that can be measured by DT-based methods. Accordingly, instead of analysing each frame independently, the temporal dynamic of each facial component is modelled to recognize specific patterns.

The system introduced by Zhao and Pietikäinen (2007) was one of the earliest attempts to use a DT-based method to recognize six basic emotions. They proposed two methods based on Local Binary Patterns (LBP). The first method tried to extend the LBP operator that was typically used for modelling static textures. In their first method, Zhao and Pietikäinen (2007) adopted the LBP operator in 3D space (x, y and time) called Volume LBP (VLBP). The VLBP method was computationally expensive and could not be used in for real-time applications. Consequently, they proposed LBP in Three Orthogonal Panels (LBPTOP), which was simpler and easier to extend. Both methods have shown reasonable accuracies for detecting the six basic emotions when applied to the images collected from CK dataset (see Table 2-3). However, Zhao and Pietikäinen (2007) needed to locate the position of the eyes manually in the first frame, which is not desirable for practical AC applications.

Another successful example of using the DT-based method for FER has been proposed by Koelstra et al. (2010). They utilized two techniques (Free-form Deformations FFDs or Motion History Images (MHIs)) to estimate non-rigid motion between consecutive frames. Instead of dividing the face region into pre-defined sub-regions, they used a quad tree decomposition method to identify specific sub-regions related to each facial AU. Then orientation histogram features were extracted from those sub-regions. They have done a comprehensive comparison between their obtained results and other current FER systems. Considering temporal features is important for discriminating those AUs that have similar appearance. For example, AU 43 (closed eyes) and AU45 (blink) are similar; however, the AU43 usually lasts longer than AU45.

Both geometric-based and appearance-based approaches have advantages and disadvantages. Each method showed better performance in certain conditions and for a group of affective states compared with another one. For example, appearance-based methods are more sensitive to light conditions, and for applications that have big variations in illumination, geometric-based methods are suggested. On the other hand, geometric-based methods are not able to measure some facial features, such as furrows and wrinkles, which appearance-based systems detected easily. There is still a debate to select the best method based on the application and targeted emotions. Some researchers (Ashraf et al., 2009; Tian et al., 2002) used both methods to extract facial features and showed that considering both types of features can improve the accuracy of affect detection.

2.5 Physiological-based affect detection

According to the theory of *Emotions as Embodiments*, several methods have been proposed to detect affective states from physiological changes. According to this theory, each internal feeling (e.g., emotion) has a specific physiological pattern. Most of the proposed methods for measuring physiological states try to record and analyse the electrical signals produced by the heart, brain, muscles, and skin. The main methods and instruments to monitor physiological signals include Electrocardiogram (ECG), Electromyogram (EMG), galvanic skin response (GSR), Skin Conductivity (SC), and Respiration (RSP). Picard et al. (2001) demonstrated the importance of using physiological signals for affect detection.

Several methods have been proposed to extract physiological features from signals recorded for affect detection. HR and heart rate variability (HRV) are two vital measures that have been used widely for affect detection and can be extracted from ECG data. Previous researchers showed that HR is a good indicator for discriminating between different affective states (Kreibig, 2010; Levenson, Ekman, & Friesen, 1990). For example, the value of measured HR during fear, anger, and sadness are higher than during happiness, disgust, and surprise (Levenson et al., 1990). Alzoubi et al. (2012) proposed a method to detect non-basic affective states using a combination of physiological signals (ECG, EMG, and GSR) during naturalistic HCI. They yielded a fair accuracy for detecting naturalistic affective using a user-independent model. However, according to the literature (Calvo & D'Mello, 2010), one of the main challenges associated with physiological-based AC applications is the *intrusiveness* of physiological sensors. This issue can be addressed by using remote measurement techniques.

2.5.1 Remote sensing of physiological signals

Remote, contactless monitoring of vital signs can be divided into three categories: microwave Doppler radar (Chen, Misra, Wang, Chuang, & Postow, 1986; Greneker, 1997; Li et al., 2009), thermal imaging (Fei & Pavlidis, 2010; Garbey, Sun, Merla, & Pavlidis, 2004), and video-based imaging methods (Poh et al., 2010; Takano & Ohta, 2007; Verkruysse, Svaasand, & Nelson, 2008). One of the earliest applications of remote contactless monitoring was reported in the 1980s (Chen et al., 1986). Chen et al. (1986) proposed a microwave life-detection system for sensing the heartbeat and breathing of human subjects lying on the ground at a distance of about 30 metres or located behind a cinder block wall. To our knowledge, this was the first effort to monitor vital signs remotely using microwave Doppler radar.

In thermal imaging, remote HR detection is performed through the analysis of skin temperature modulation. Pulsative blood flow modulates tissue temperature because of the heat exchange by convection and conduction between vessels and surrounding tissue. Such modulation is more pronounced in the vicinity of major superficial blood vessels (Garbey et al., 2004). A superficial blood vessel should be selected as the region of interest (ROI) for image analysis. Detecting and tracking these areas on human bodies is a challenging task. On the other hand, remote breathing measurements take advantage of the fact that the expired air has a higher temperature than the inspired air due to heat exchange in the lungs and respiratory passageways. This thermic nature of breath around the nostril area creates an opportunity for thermal measurement (Fei & Pavlidis, 2010). Sprager and Zazula (2013) proposed a new method to measure HR and respiration rate from optical interferometric signals under two separate conditions: at rest and during cycling.

Among the above mentioned methods, video-based remote monitoring methods are considered cheaper and easier to adopt (Poh et al., 2011). Most of them use the photoplethysmography (PPG) methodology to detect HR variability. PPG is a low-cost and non-invasive means of sensing the cardiovascular blood volume pulse through variations in transmitted or reflected light (Allen, 2007). PPG is based on the principle that blood absorbs light more than the surrounding tissue so variations in blood volume affect transmission or reflectance correspondingly.

PPG is measured through reflection of dedicated light sources, such as infrared, with a shallower penetration depth in skin. Researchers (Takano & Ohta, 2007; Verkruysse et al., 2008) demonstrated that pulse measurement could be obtained by analyzing the skin colour with ambient light as the illumination source. Poh et al. (2011) proposed the method for recovering BVP signals by tracking the face skin colour changes using an ordinary camera. It has shown its robustness under a non-laboratory condition with little user motion. Automatic face tracking ability enhances the method to work well in a wide range of everyday activities.

2.6 Multimodal affect detection

It is generally accepted that integrating information from multiple cues can improve the performance of affect detection. However, a multimodal affect detection system has been rarely explored before due to difficulties associated with recording and synchronization of multiple sensors (Jaimes & Sebe, 2007). The integration of multiple cues can be implemented in three levels. Figure 2-5 illustrates these three levels of fusion. In *Data-level fusion* (Figure 2-5-a), the raw data from all sensors is integrated in the early stage before feature extraction and classification. Recorded data must have the same temporal resolution to combine them in data-level. In addition, this type of fusion cannot be used for combining non-homogenous modalities. For example, video and text data cannot be integrated in data-level. Because of these mentioned limitations, it has been rarely used for affect detection.

However, *Feature-level fusion* goes one step further and combines extracted features from each modality. With this approach, non-homogonous modalities with different recorded frequencies can be integrated. Feature-level fusion is more common in AC applications (Busso et al., 2004; Schuller et al., 2007; Wagner, Kim, & André, 2005). After combining all extracted features, a single or a meta classifier is applied on the data set for affect prediction (Figure 2-5-b).

Decision-level fusion is also a popular approach for multimodal affect detection (D'Mello & Graesser, 2010; Pal, Iyer, & Yantorno, 2006; Z. Zeng et al., 2007). In this approach, the integration occurs at the end of the analysis (Figure 2-5-b). To be more precise, the prediction results from each modality are combined to conclude the final decision. It was assumed that the decision-level approach can provide better recognition rate; however, several experiments showed that the feature-level approach can outperform the late-integration approach (Castellano, Kessous, & Caridakis, 2008; Kim, 2007).

For example, Kim (2007) evaluated these two approaches for detecting dimensional affect using speech and physiological changes. The feature level fusion obtained better accuracies compared with other fusion approaches. Choosing an appropriate approach to integrate different modalities depends on the type and number of modalities, context, and application. Recently the *Hybrid-fusion* approach (Hussain et al., 2011; Kim, 2007) was introduced and this tries automatically to select the best solution by combining feature level and decision level integration.

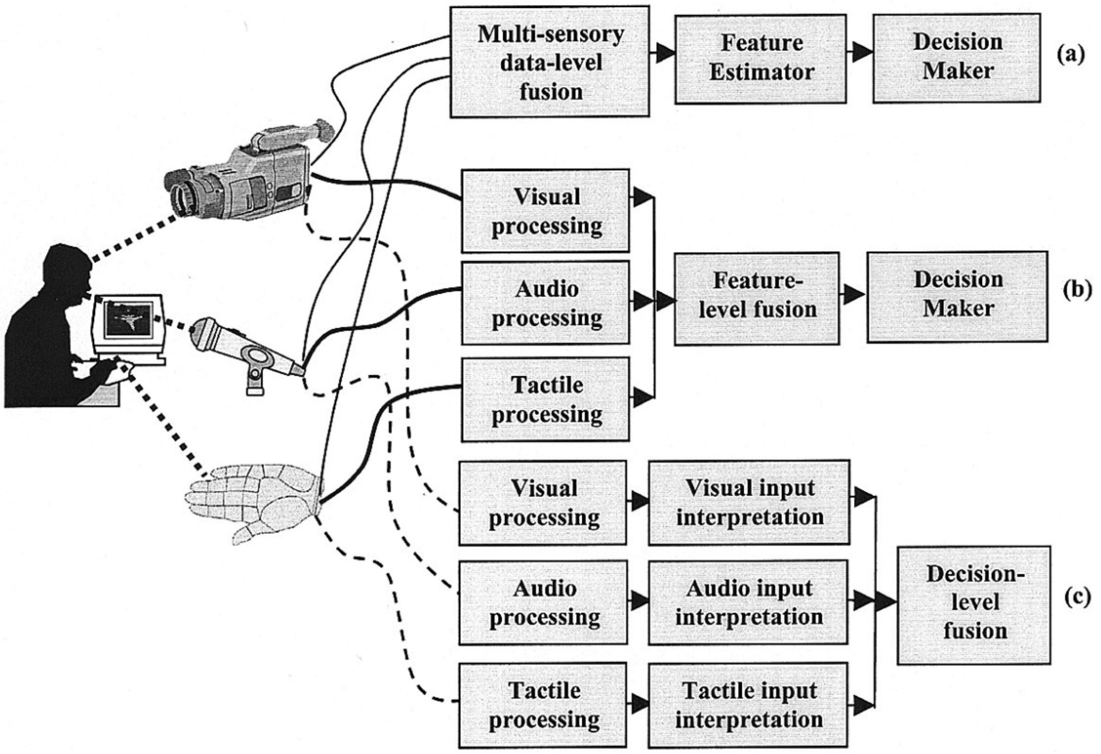


Figure 2-5: Three levels of fusion in multimodal system for affect detection. From Pantic and Rothkrantz (2003)

Recent attempts of multisensory AC systems focused on detecting non-basic emotions (Caridakis, Karpouzis, & Kollias, 2008; Nicolaou, Gunes, & Pantic, 2011; Soleymani, Pantic, et al., 2012). Combining audio-visual features received more attention compared with others (Zhihong Zeng et al., 2009). Combinations of other modalities, such as text and audio

(Forbes-Riley & Litman 2011), tactile and physiological signals, and face and physiological signals, were also explored.

Facial features were usually combined with other audiovisual modalities like head and upper-body gestures or audio cues, such as acoustic and prosodic features. On the other hand, physiological responses were mostly fused with each other, as long as all features extracted from wearable and tactile sensors (Alzoubi et al., 2012; Chanel, Ansari-Asl, & Pun, 2007; Wagner et al., 2005). Facial expressions and physiological signals were rarely explored together (Bailenson et al., 2008; Hussain, Calvo, & Chen, 2013).

To descriminate two types of emotion (amusement vs. sadness), Bailenson et al. (2008) presented a system to monitor 15 physiological measures and 53 facial points when participants were watching a film. Bailenson et al. (2008) combined the data at feature level and evaluated two classifier (SVM and GentleBoost) for emotion classification. The best F1-score (0.69) was achieved by the fusion of the face and physiology for amusement detection. Their evaluations also showed that the fusion model improved the F1-score of the face and physology modalities by 0.06 and 0.20, respectively.

Recently, some works focused on continuous prediction of affective states using audiovisual information in dimensional space (Gunes & Schuller, 2013; Nicolaou et al., 2011). Nicolaou et al. (2011) fused facial expression, shoulder gestures, and audio cues to predict affect in 2D space (valence and arousal) continuously. They applied two machine learning techniques: Bidirectional Long Short-Term Memory neural networks (BLSTM-NNs) and Support Vector Machines for Regression (SVR) on extracted features and evaluated the performance of each ones. The results showed a significant agreement between system predictions and human coders annotations. They validated their system using the leave-one-sequence-out cross validation approach. A root mean squared error (RMSE) of 0.15 and a correlation of 0.796

were obtained for valence prediction, and a RMSE of 0.21 and correlation of 0.642 were achieved for arousal prediction.

Chapter 3. Image analysis and classification methods

Summary

In this chapter, the proposed framework for recognizing non-basic affective states by fusing three types of vision-based features is described in detail. Two main experiments were designed and conducted to elicit facial expression in a naturalistic environment and are described here. In the first experiment, a set of emotional images selected from the IAPS collection was used to trigger a user's emotions. For the second study, a writing session was designed to monitor user's engagement and emotion during the writing task.

In addition, the methods used for feature extraction and for recognizing expressions are presented in the rest of this chapter. Three methods are introduced for feature extraction. First, the LBPTOP is a dynamic texture-based method to extract appearance-based and motion-based features. Second, the Microsoft Kinect face tracker engine (SDK v. 1.5) was used to extract geometric-based features. Third, the HR signals recovered by computer vision techniques were also used for affect detection. A voting classifier is used for classification, and the Cohen Kappa measure was used to measure the performance of the proposed methods.

3.1 Introduction

The literature review of Chapter 2 showed a gap in our knowledge on designing and implementing automatic non-basic affective state recognition in naturalistic scenarios (D'Mello & Calvo, 2013; M. E. Hoque et al., 2012). The literature also suggested that considering the geometric appearance of the face provides a better understanding of the affective state (Ashraf et al., 2009; Tian et al., 2002). Adding other modalities, like physiological signals, can improve the affect prediction by providing some useful information about the user's intrinsic feelings. We present a framework for recognizing non-basic affective states to combine those three mentioned factors: appearance- and geometric-based features and physiological signals. Our proposed framework relies only on vision modality. With new advances on computer vision techniques, some physiological signals could be extracted through video recording. Gathering more useful information using just one modality (e.g., video) is less intrusive and easy to synchronize. A comprehensive overview of our proposed framework and its components is presented in this chapter.

Basically, the affect recognition process consists of two main steps (Fasel & Luettin, 2003): feature extraction and classification. In this chapter, two different methods are introduced for extracting appearance-based and geometric-based features from facial videos. In addition, an improved method for recovering HR signals is introduced to extract some physiological features. A combination of these features is passed to the classification component to detect the corresponding affective state. A number of classifiers is introduced and evaluated in this chapter.

Existing corpora of video-based facial expressions are focused mostly on basic emotions and recorded in a controlled conditions (M. E. Hoque et al., 2012). Most of them used actors to play a facial expression. Quite a few accessible corpus still contain non-basic affective states expressed by ordinary people in a naturalistic scenarios (Mckeown et al., 2010). Non-basic affective states rarely happen in everyday activities. Gathering these kinds of expressions and labelling them is a tough and time-consuming task. Reliability of gathered data is another issue, which is more important for these sorts of corpora. We have decided to build a new corpus of naturalistic affective states to evaluate our proposed framework. This corpus is also described in this chapter.

3.2 Framework overview

The main objective of the proposed framework is handling disparate video-based features for recognizing non-basic affective states. Facial expressions and physiology are two modalities that could be monitored by computer vision techniques. Figure 3-1 illustrates an overview of our methodology. Raw data from video and depth sensor composes the input of the system. Three types of features are extracted using three different methods. Facial AUs and head-related features are extracted as a type of geometric-based features. The LBPTOP features are the second type of features that represent the appearance-based features. The third type of features is extracted from the HR channel. A combination of these features is used for affect classification.

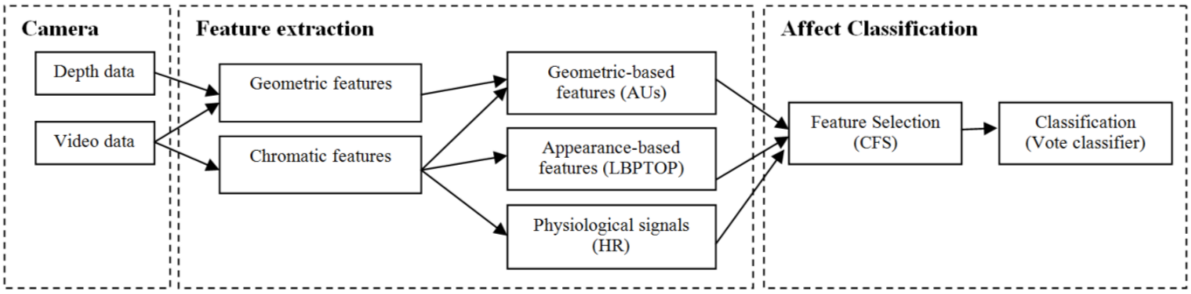


Figure 3-1: The proposed model for classifying naturalistic complex mental states

The fundamental criteria for implementing this framework are:

- Dynamic representation of affect: Frames inside each video that can be analysed independently as static images are not enough for an accurate estimation of an affective state. To have a better understanding of a user's affective state, a temporal model of affect should be considered. This fact is one of the most important factors in designing feature extraction methods in the proposed framework. In particular, appearance-based features are extracted based on a dynamic-texture method.
- Unobtrusiveness: In general, contact-less sensors are less-intrusive compared with wearable sensors, such as current physiological sensors. New generation of cameras could observe some vital signs, such as HR and respiration, remotely. The main concern about using a camera as a sensor is privacy. Some people are not comfortable with systems that film them as they consider them a threat to their privacy. It is fair to consider a trade-off between privacy and performance. If the system could provide a reasonable level of performance, the user could be conceived to share more private data. On the other hand, current advances technologies for enhancing privacy could decrease the privacy concerns in the near future (e.g., local processing could mean that only anonymised features are sent outside the device or all the processing happens locally).
- Synchronization: In multimodal affect detection systems, synchronization has always been a challenging issue. Different sensors have different frequencies and might use different time services. Having one sensor could reduce the complexity of synchronization task. The proposed framework relies only on a visionary sensor (Microsoft Kinect) that can extract facial features and physiological signals.

3.2.1 Geometric-based features

Geometric-based features can be extracted by detecting certain points of the object (e.g., face) in an image and tracking them through a sequence of images. Several methods and techniques have been proposed following this approach as discussed in Chapter 2. Among all of the proposed methods, AUs (Ekman, Friesen, & Hager, 2002) are still one of the best-known and useful methods to describe facial expressions. Recent methods are able to detect a combination of AUs automatically in real time (Baltrusaitis et al., 2011; Koelstra et al., 2010; Torre, Simon, Ambadar, & Cohn, 2011). For instance, Microsoft introduced a software development kit (SDK) for its Kinect sensor with the ability to detect 100 facial points and 6 AUs in real time. This SDK was used in this thesis to extract geometric-based features.

3.2.2 Appearance-based features

Appearance-based features aim to describe deformations of facial objects, such as wrinkles and furrows, which geometric parameters find hard to model. LBP (Ojala, Pietikäinen, & Harwood, 1996) is a powerful method used for detecting micro-patterns in an image. This method is also used for FER systems (Shan et al., 2009). LBP itself could not support dynamic representation of facial expressions, but LBP in three orthogonal planes is proposed to extend the LBP into temporal space.

3.2.3 Remote sensing of physiological signals

Currently, two important vital signs – heart pulse and respiration – could be extracted using ordinary cameras (Poh et al., 2011). There are still some difficulties with accurate estimation of these signals due to lighting conditions and the user’s body movement. Despite these issues, remote monitoring of physiological signals could improve our understanding about the user’s affective states. Researchers have demonstrated that physiological signals are a useful indicator of non-basic affective states (Alzoubi et al., 2012). In the proposed framework, a component is devoted to remote sensing of heart pulses.

3.3 Feature extraction

3.3.1 Kinect face tracker

In this study, the Kinect SDK’s face tracking engine (v1.5) was used for facial feature extraction. This engine is able to track head position, ANimation Units (ANUs) and 100 facial points in real time. It should be mentioned that the description of these ANU labels are different from typical AUs proposed by Ekman and Friesen (1978). For example, ANU0 (Upper Lip Raiser) equals to AU10 in the FACS proposed by Ekman and Friesen (1978). A detailed description of each ANU is presented in Table 3-1. Six ANUs are tracked by the face tracking engine and these are a subset of what is defined in the Candide3 model (Ahlberg, 2001). The ANUs are deltas from the neutral face shape. Four ANUs represent lips motions and two correspond to eyebrows motions. Each ANU is expressed as a numeric weight varying between –1 and +1. For example, ANU0 measures Upper Lip Raiser. The value of 0 means the upper lip covers the teeth fully; ANU0 equals +1 if the user shows his/her teeth fully; and it decreases toward –1 if he/she pushes down the lip.

Table 3.1 Animation unit (ANUs) description according to Microsoft Kinect SDK 1.5 user manual

ANU Name and Value	ANU Value Interpretation
ANU0 – Upper Lip Raiser (Equals to AU10)	0 = neutral, covering teeth 1 = showing teeth fully –1 = maximal possible pushed down lip
ANU1 – Jaw Lowerer (Equals to AU26/27)	0 = closed 1 = fully open –1 = closed, like 0
ANU2 – Lip Stretcher (Equals to AU20)	0 = neutral 1 = fully stretched (joker's smile) –0.5 = rounded (pout) –1 = fully rounded (kissing mouth)
ANU3 – Brow Lowerer (Equals to AU4)	0 = neutral –1 = raised almost all the way +1 = fully lowered (to the limit of the eyes)
ANU4 – Lip Corner Depressor (Equals to AU13/15)	0 = neutral –1 = very happy smile +1 = very sad frown
ANU5 – Outer Brow Raiser (Equals to AU2)	0 = neutral –1 = fully lowered as a very sad face +1 = raised as in an expression of deep surprise

The coordinate system used for measuring the position and rotation of the head is based on a right-handed coordinate system, and the origin of the coordinate is the position of the Kinect sensor. The Z axis is pointing toward the user, the X axis is pointing toward the right side, and the Y axis is pointing up.

The user's head rotations are captured by three parameters in 3D space: head *pitch*, *yaw*, and *roll*. The head pitch angle is measured in degrees and specifies the rotation of the head around the X axis. To be more precise, it can measure if the user is looking down or up. The head yaw angle can measure the rotation of the head around the Y axis. If the user is looking toward his/her right shoulder, the value of yaw angle would be negative, and if he/she is turning his/her head toward his/her left shoulder, this value would be positive. The head roll angle measures the rotation of the head around the Z axis. Figure 3-2 shows the description of these features graphically.

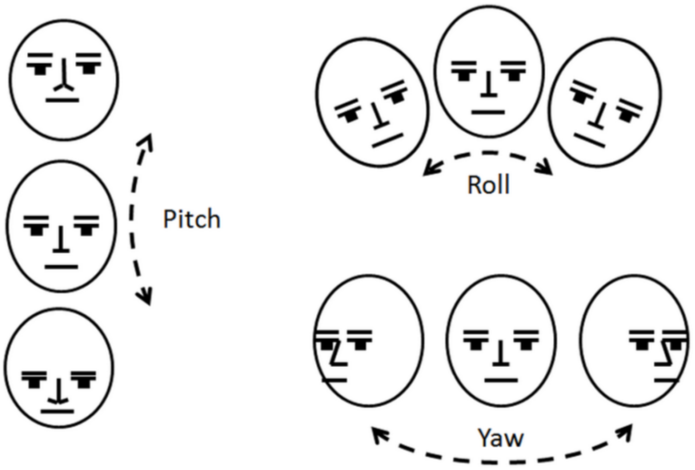


Figure 3-2: Head rotation measurement examples from Microsoft Kinect SDK manual [1]

According to the origin of the coordinates, the head position is also measured in meters in 3D space using the head *translation* features in the X, Y, and Z axes. Six ANUs were calculated for each detected face, along with three values that specify the head rotation in 3D space and three values that indicate the position of the head. Accordingly, 12 features were calculated for each frame. To add temporal features, some statistical functions are applied on extracted features. Considering each video segment as an instance, seven statistical functions (mean, median, standard deviation, max, min, range, difference) were applied on each feature to build the final set of 84 (12 features × 7 statistical functions) features.

3.3.1.1 *Validation*

Lighting conditions, distance, and face occlusion are the main factors that might affect the accuracy of the face-tracking engine. For example, thick glasses or certain facial hair might reduce the accuracy facial point and the ANU tracking. The face-tracking engine is also sensitive to the position of the head in front of the camera. According to the SDK user

[1] http://msdn.microsoft.com/en-us/library/jj130970.aspx

manual[2], it can track well when the user's head pitch angle is less than 20 degrees, the roll angle is less than 90 degrees, and the yaw angle is less than 45 degrees. But it works best when the pitch angle is less than 10 degrees, the roll angle is less than 45 degrees, and the yaw angle is less than 30 degrees.

The Kinect sensor works in two modes: *default* and *seated* mode. In SDK 1.5, if the sensor is set to *seated* mode, the application can receive full joint information when tracking users as close to the sensor as 0.4 meters up to a maximum of 3.0 meters. The user tracking range in the default mode is from 0.8 meters to a maximum of 4.0 meters. The seated mode is more appropriate for normal HCI applications. According to these limitations and to remove low quality and invalid records from extracted features, constraints were applied on extracted features (see Table 3-2).

Table 3.2: Constraints for head movements that used in the proposed framework

	Constraints used in proposed framework		Constraints in the SDK 1.5	
	Minimum value	Maximum value	Best performance	Tracking range
Distance	0.6 meter	1.4 meter	–	0.4 to 3.0 meters (in near mode)
Pitch	– 20 degrees	+ 20 degrees	± 10 degrees	± 20 degrees
Roll	– 90 degrees	+ 90 degrees	± 45 degrees	± 90 degrees
Yaw	– 45 degrees	+ 45 degrees	± 30 degrees	± 45 degrees

3.3.2 Local Binary Pattern in Three Orthogonal Planes

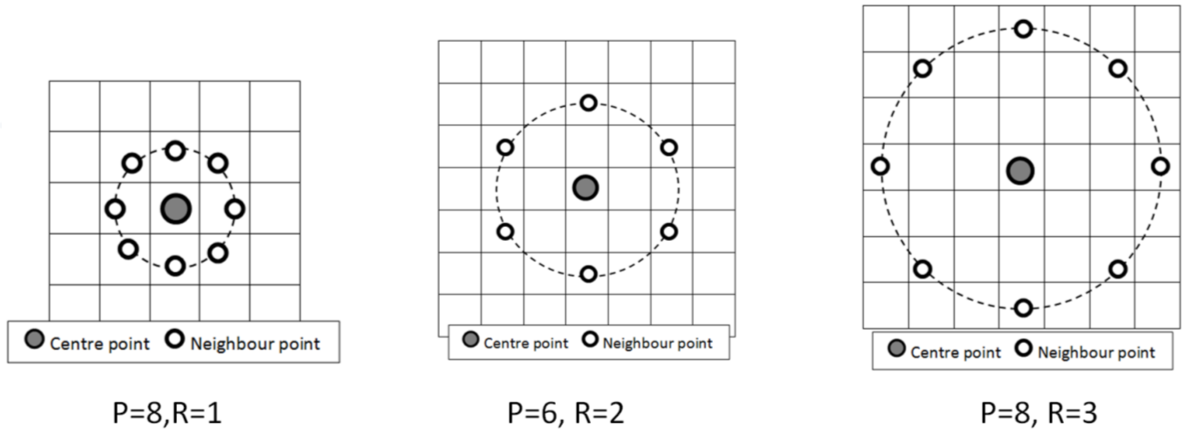
3.3.2.1 *Local Binary Pattern (LBP)*

The LBP method detects the local-patterns existing in an image. Ojala et al. (1996) proposed this method for describing texture images. By applying the LBP operator on all pixels of an image (or a sub-region of an image) and computing the distribution of local-patterns, a unique histogram could be extracted that describes the occurrence of each specific local

[2] http://msdn.microsoft.com/en-us/library/jj130970.aspx

pattern throughout that image. This histogram is a powerful identifier for each image and shows good performance in several pattern recognition applications.

For detecting each local-pattern, the LBP operator obtains the colour value of each centre point and considers it as a threshold value. Then it compares the colour value of neighbourhood pixel with the threshold and assigns a binary value to that pixel as its label. The LBP operator puts labels of the neighbouring pixels together to come up with a binary number that specifies a unique local-pattern. Having P neighbourhood pixels could describe 2^P distinguishable local-patterns. For example, if we consider eight neighbourhood pixels, 256 local-patterns could be addressed. The number of neighbourhood pixels (P) and the radii (R) could define the size of each local-pattern. The radii specify the distance between neighbourhood pixels and the centre point. The ideal values for R and P depend on the application domain and the characteristics of the image. Three examples of LBP operators with different values for R and P are presented in Figure 3-3.


● Centre point ○ Neighbour point
● Centre point ○ Neighbour point
● Centre point ○ Neighbour point

P=8,R=1

P=6, R=2

P=8, R=3

Figure 3-3: Examples of different types of LBP operators

Figure 3-4 illustrates the procedure of calculating a local-pattern using a LBP operator (P = 8, R = 1). It is looking for a local-pattern in a 3 × 3 pixel block. According to the colour values, the threshold value was set to 120. The labels of neighbourhood pixels were calculated based on the threshold value. Finally a binary number was created by putting those labels together based on the clockwise circular order, which indicates that an instance of *local-pattern #184* was detected using this procedure.

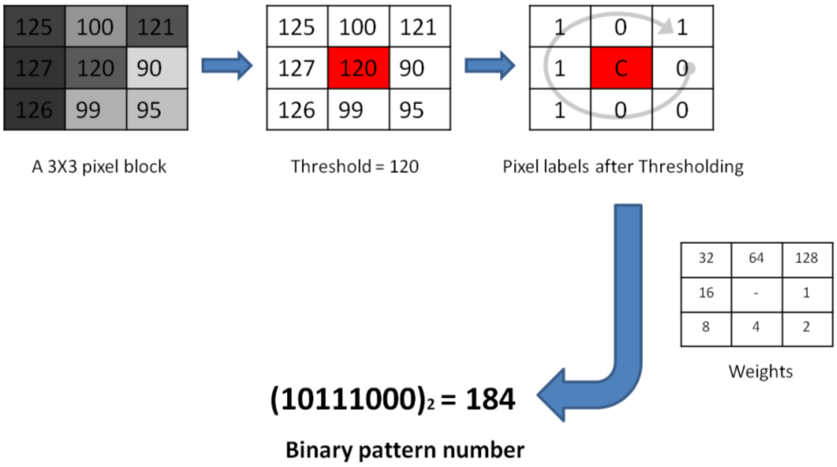
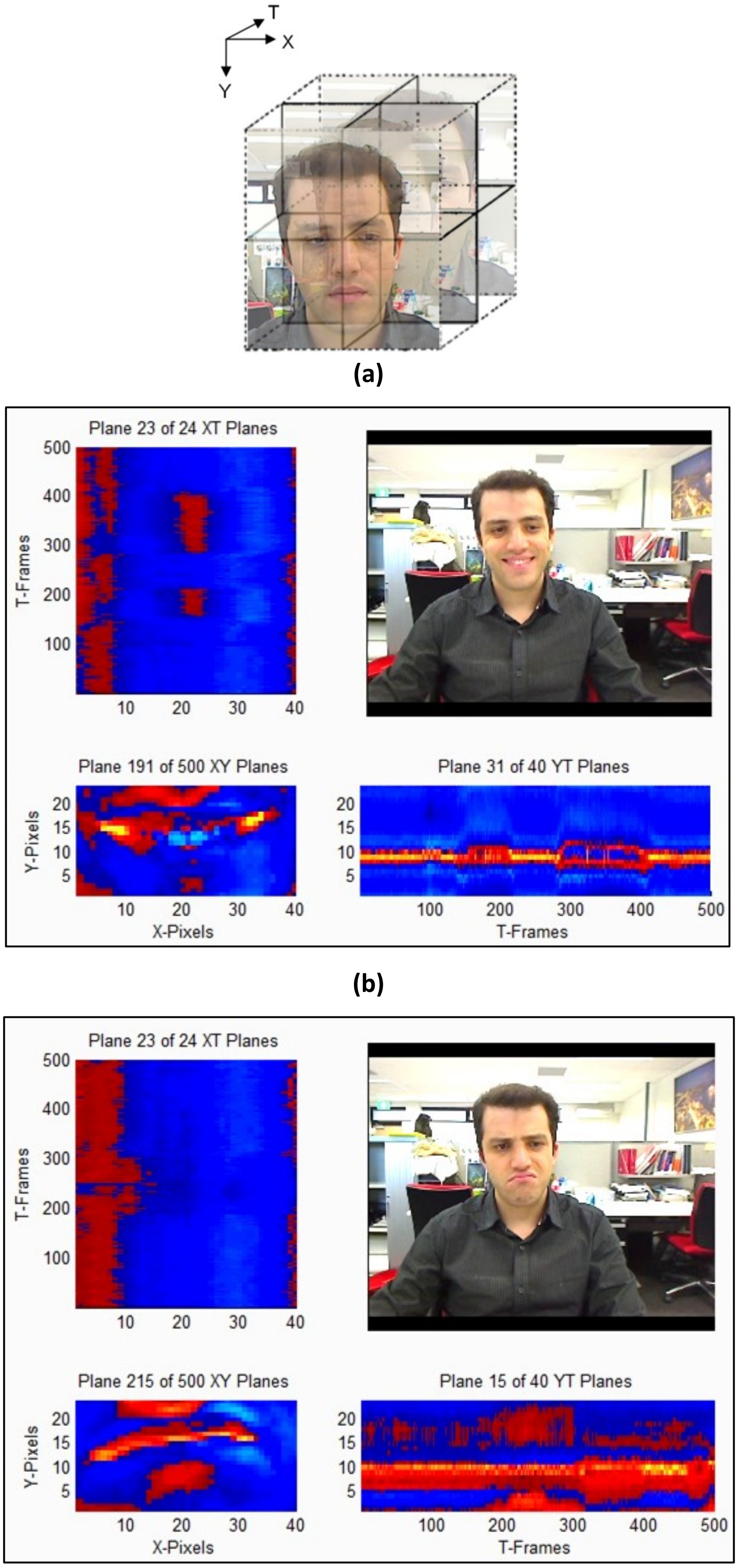


Figure 3-4: The procedure of calculating LBP when P = 8, R = 1

3.3.2.2 *LBP in Three Orthogonal Planes (LBPTOP)*

LBP methods were originally proposed for recognizing static texture images. Different variations of this method have been proposed for recognizing dynamic textures in different applications. Zhao and Pietikäinen (2007) introduced one of the most successful extensions of LBP for detecting facial expressions. To extract LBP features from a video segment, the authors divided a video segment into three sets of orthogonal planes. A video segment could be considered as a sequence of static images (XY planes) in the time axis. In another perspective, we can analyse a video segment as a stack of XT planes in the Y axis or a stack YT planes in the X axis (see Figure 3-5). Figure 3-5-a shows that how a video segment is divided into three sets of orthogonal planes. Figures 3-5-b and 3-5-c illustrate an instant from each set of planes extracted from mouth area for two for two different affective states (happiness and disgust).



(a)

(b)

(c)

Figure 3-5: (a) Dividing a video segment into three orthogonal planes. (b) Happiness example. (c) Disgust example

Zhao and Pietikäinen (2007) calculated the LBP for each set of planes and created three separate histograms for each set. Detected local-patterns in the XY planes and local-patterns in XT and the YT planes represent the dynamic motions of each local region of the image in the Y and X directions, respectively. By concatenating these three histograms, the LBPTOP features were created for a video segment. Accordingly, 3×2^P features could be extracted for each video segment (where P = number of neighbouring pixels). However, the radius in axes X, Y, and T and the number of neighbouring pixels in the XY, XT, and YT planes can also be different, and can be marked as R_X, R_Y, and R_T, P_{XY}, P_{XT} , and P_{YT}. The corresponding set of features is denoted as $LBPTOP_{PXY;PXT;PYT;RX;RY;RT}$. In this thesis, three variations of the LBPTOP method were used for feature extraction that were different in the radius and in the number of neighbourhood points ($LBPTOP_{8,8,8,1,1,1}$, $LBPTOP_{6,6,6,2,2,2}$ and $LBPTOP_{8,8,8,3,3,3}$).

Each LBPTOP description provides information about the numbers of occurrences of each specific local-pattern throughout the video segment without any extra information about their occurrence locations. This is not desirable for some applications like FER systems. In FER, a specific local-pattern in one region of the face (e.g., eyes) has a different meaning from the same local-pattern in another region of the face (e.g., mouth). Accordingly, Zhao and Pietikäinen (2007) proposed to divide the face area into several blocks and calculate LBPTOP histograms for each block separately. Finally, by concatenating all histograms, a new set of features could be obtained that considers the approximate location of each local-pattern. If the image divides into N blocks, the number of bins (features) in the final histogram would be $N \times 3 \times 2^P$.

In our proposed model, the face is tracked using an extended boosted cascade classifier (Viola & Jones, 2001) implemented in Open source Computer Vision (OpenCV[3]) library.

[3] http://opencv.org/

Then three blocks of facial components are extracted from the detected face region: left-eye, right-eye, and mouth. To have the same size of blocks in each image, the detected objects are resized to fixed sizes. The LBPTOP operator is applied on each block separately, and the final feature set is created by concatenating the results from each block. Totally, $3 \times 3 \times 2^P$ features are extracted from each video segment.

3.3.3 Video-based HR measurement

Among the current methods of remote physiological sensing, video-based techniques are considered cheaper and easier to adopt. Most of these attempts use the PPG methodology to detect cardiovascular blood volume pulse. The PPG method was introduced in 1937 (Hertzman & Spealman, 1937) and typically measures the light reflection from the skin tissue to detect physiological signals like oxygen saturation (pulse oxymetry), HR, and blood pressure. It works based on the principle that the amount of light absorption differs significantly between blood loaded with oxygen and blood lacking oxygen. Accordingly, after each heart pulse, fresh blood is pumped to the skin and the capability of light absorption will increase. To be more precise, variations in blood volumes in the skin tissue can affect the transmission or reflectance. As shown in Figure 3-6, detectable changes in light absorption caused by heart pulses (pulsatile) are very small compared with the absorption due to non-pulsatile arterial blood. These small changes should be detected and amplified to measure HR signal.

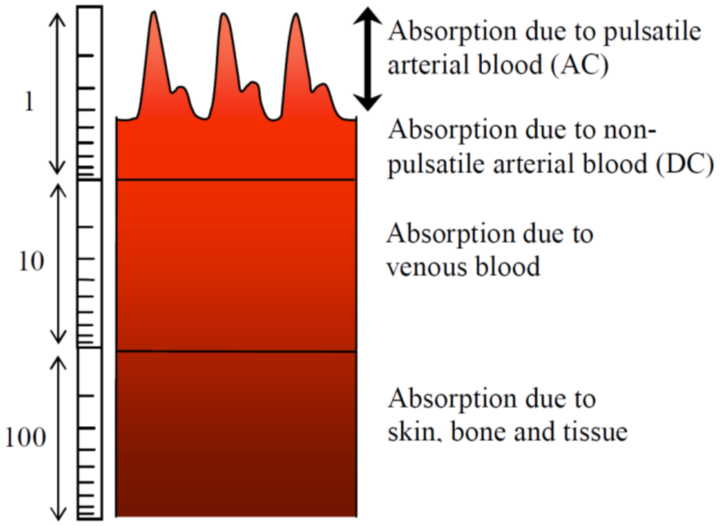

1
10
100
Absorption due to pulsatile arterial blood (AC)
Absorption due to non-pulsatile arterial blood (DC)
Absorption due to venous blood
Absorption due to skin, bone and tissue

Figure 3-6 A breakdown of light absorbtion in the PPG signal adopted from Cheang and Smith (2003)

Typically, PPG uses a dedicated light source that emits light (e.g., infrared wavelength) into the skin and measures the transmission or reflectance using another sensor. To the best of our knowledge, almost all PPG sensors need to be in contact with the skin to avoid ambient lights that might introduce noise (Cheang & Smith, 2003; Hummler, Engelmann, Pohlandt, Högel, & Franz, 2004). However, Verkruysse et al. (2008) showed for the first time that PPG signals could be detected remotely on the human face with a normal ambient light (as the only illumination source) using an ordinary digital camera. This attempt opened a door to the multiple new applications in the remote physiological sensing research area. Later, Poh et al. (2010) introduced an improvement on their work, using the Independent Component Analysis (ICA) method. They compared their algorithm with a BVP sensor and achieved a Pearson correlation coefficient of 0.98 for detecting HR at rest.

We have proposed a new method (described in (Monkaresi, Calvo, & Yan, 2014)) to extract HRs from facial video recording based on the algorithm proposed by Poh et al. (2010). Figure 3-7(a) shows the flowchart of automatic recovering HR signals as described by Poh et al.

(2010). Our variation of their method is also illustrated in Figure 3-7(b). Different steps of the proposed method are explained.

3.3.3.1 *Face Tracking*

The goal is recovering the heart pulses from the face. The first step is to detect and track the face in the recorded or live video. An extended boosted cascade classifier implemented in OpenCV (v. 2.2) library is used for face tracking (Viola & Jones, 2001). The PPG signal could not be recovered from some parts of the face covered by hair. The algorithm should focus on the regions more likely contain uncovered skin like forehead and cheeks. On the other hand, the face area detected by OpenCV library might contain parts of the background regions. These regions should be omitted before further analysis. Poh et al. (2010) suggested that we could consider the centre 60% of width of the detected face as the ROI to make sure that there are no unwanted background regions.

In addition, to increase the calculation speed and reduce the false face detection rate due to background artefacts (false positives), a robust face tracking algorithm is implemented. According to this method, after detecting the face for the first time, the algorithm remembers the location of the face in the image for the next frame and the next face detection task starts by focusing on that region. If the face were not found, the search region is expanded to the whole image and this procedure is repeated for the next frames.

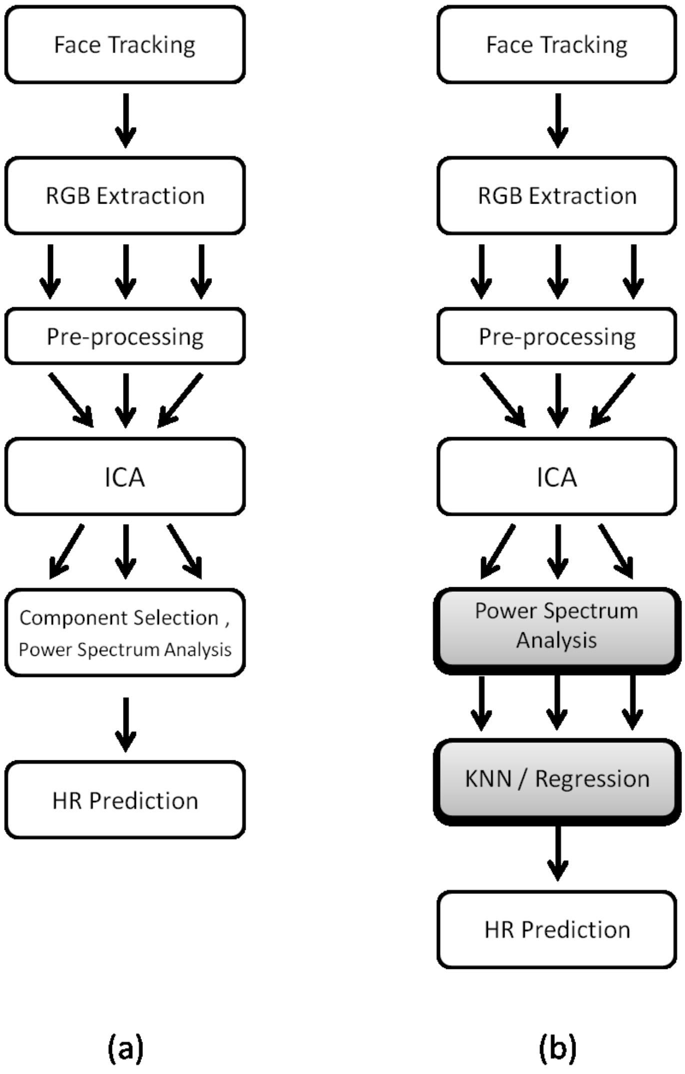


Figure 3-7: The flowchart of HR extraction from video recording. (a) The method proposed by Poh et al. (2010); (b) Extended method improved by machine learning techniques

3.3.3.2 *RGB Extraction*

To recover the HR signal, a sequence of images needs to be considered. The length of each sequence is set to 30 seconds. As mentioned in the face-tracking sub-section (Section 3.3.3.1), a rectangular area of 60% width and the full height of the detected face is considered as the ROI in each frame. Each ROI is divided to the RGB channels and the average of each colour (RGB) amplitude value is calculated across all pixels in the ROI. The sequence of these averages based on the length of the measurement (30 or 60 seconds) composes the raw signals for red, green, and blue channels. These three raw signals are considered as the inputs for the ICA.

3.3.3.3 *Pre-processing*

Before applying ICA, the raw traces are deterended and normalised to improve the quality of the signals. Deterending could remove the long-term fluctuation in the baseline of the RGB traces that might occur by fast head movements. A smoothness priors implementation (Tarvainen, Ranta-aho, & Karjalainen, 2002) is used for deterending the data. Then, each deterended signal $x(\mathrm{t})$ is normalized based on the following equation (Equation 3-1):

$$\boldsymbol{x}'(\boldsymbol{t}) = \frac{\boldsymbol{x(t)-\mu}}{\boldsymbol{\sigma}} \qquad \textbf{3-1}$$

where μ and σ are the mean and standard deviation of $\mathrm{x(t)}$, respectively. The ICA performs on these normalized raw traces.

3.3.3.4 *Independent Component Analysis (ICA)*

ICA (Comon, 1994) is a special case of the Blind Source Separation (BSS) techniques, which try to separate a multivariate signal into statistically independent subcomponents by assuming that the subcomponents are non-Gaussian signals. To be more precise, ICA finds the independent components by maximizing the statistical independence of the estimated components. To do this, two main approaches were proposed:

1. Minimization of mutual information by measuring statistical factors like maximum entropy.

2. Maximization of non-Gaussianity using iterative methods to minimize the cost functions like kurtosis and negentropy that are used to measure non-Gaussianity. The central limit theorem motivated these sort of algorithms (Trotter, 1959).

Here we adopted a linear ICA based on the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm (Cardoso, 1999). In the linear ICA, it is assumed that the observed signals contain linear mixtures of source signals. Typically, the ICA cannot identify the actual number of source signals but the number of recoverable sources is less than or equal to the number of observations. So the observed data are represented by the random vector $x = (x_1, x_2, x_3)^T$ and the source components as the random vector $s = (s_1, s_2, s_3)^T$. The goal is finding a matrix W to transform the observed data x into independent components s as shown in Equation 3-2.

$$\boldsymbol{s} = \boldsymbol{W}\,\boldsymbol{x} \tag{3-2}$$

To do this, the ICA assumes that each observed signal x_i is defined as a linear mixture of source signals s_k, $(here\text{: } k = 1, 2, 3)$, as shown in Equation 3-3:

$$\boldsymbol{x_i} = \textstyle\sum_{j=1}^{3} \boldsymbol{a_{ij}}\ \boldsymbol{s_j} \tag{3-3}$$

where the a_{ij} are the mixing weights for each x_i. Equation 3-3 can be represented in a vertical and compact format:

$$\boldsymbol{x} = \boldsymbol{A}\,\boldsymbol{s} \tag{3-4}$$

where the mixing matrix $A = (a_1, a_2, a_3)$ consists of coefficient columns $a_i = (a_{i,1}\,, a_{i,2}\,, a_{i,3})\,, i = 1,2,3$.

The task is estimating the mixing matrix A and source signals $\boldsymbol{s}$ using this model and observation values (x_i vectors). A cost function, which either maximizes the nongaussianity of the calculated $s_i = w^T x$ or minimizes the mutual information, needs to be set up. This could be done using an iterative approach to maximise or minimise the cost function. In fact,

W is an estimation of the inverse mixing matrix A and the original sources could be recovered by multiplying the observed signals x with the unmixing matrix (Equation 4-2).

3.3.3.5 *Component selection*

The ICA is also unable to determine the correct ordering of the source signals. So to identify the component that contains the HR signals further, analysis need to be done. Poh et al. (2010) selected the second component manually, as they argued that the HR signal could be observed clearly from that component. They computed the Power Spectrum Density (PSD) curve for the second component and considered the frequency of highest peak in the PSD curve as the frequency of heart beats. However in their latest report (Poh et al., 2011), they were looking for the HR signal among all three components. For this reason, they performed the PSD analysis on the three output components, and then the component that contained the spectrum with the highest peak among all spectra was selected. The frequency of that spectrum was considered as the frequency of the cardiovascular signal.

The later approach (Poh et al., 2011) is more systematic and reliable than the manual component selection. It was assumed that the HR signal is the most powerful spectrum in the operational frequency range. The operational range was set to [0.75, 4] Hz corresponding to [45, 240] bpm to provide a wide range of HR measurements. However, other sources might provide some noises in this operational range that are more powerful than the HR signal. In these cases, the highest peak in the PSD curve does not represent the HR signal. The HR signal might be represented through other peaks (for example, the second peak) in the same component or even in the other components. Poh et al. (2010) proposed a noise reduction method to address this issue.

Besides those estimations, the spectrum which contains the highest peak among all three components can be considered as the third estimation. We refer to this estimation as MPA

(Maximum Peak among All) in the rest of this dissertation. The component selection module has been improved by ML techniques in our implementation.

3.3.3.6 *Noise reduction*

To avoid selecting wrong values produced because of noises, Poh et al. (2010) proposed a historical estimation method. This method set a threshold of 12 bpm and evaluated the current estimation of the HR with the previous accepted estimation (taken 1 second apart). If the absolute difference between current and previous estimation was less than or equal to the threshold, the current estimation was accepted. Otherwise, the current estimation was rejected and the method looked for HR frequency by evaluating the next highest peak in the operational frequency range that met this constraint. If there was no peak in that range that met the constraint, the last accepted estimation was considered for the current estimation.

3.3.3.7 *Machine learning techniques*

We proposed a new technique to find the spectrum that contains the HR signal. We cannot rely only on one component to obtain the HR signal, particularly on noisy occasions when the observed signals are affected by head movements and changes in illumination. By obtaining some information from each output components and applying ML techniques, a better estimation could be obtained. In our proposed method, nine features were extracted from the three PSD curves of the independent components. The set of features includes the frequency of highest peaks in the PSD curves before and after applying noise reduction method and the depth of searches in the noise reduction method for each component to form the rest of the features. Two machine learning techniques – Linear Regression and k-Nearest Neighbour (kNN) – are proposed for HR estimation using these nine features.

Linear regression is a method that tries to model a linear relationship between one dependent variable and one or more independent variables by fitting a linear equation to the observed data. Here, the HR is the dependent variable and the extracted features are independent variables. This linear equation describes how the heart rate (hr) changes with nine explanatory variables ($\mathrm{f_{ij}}, \mathrm{j} = 1,2,...,9$). Accordingly, the multiple linear regression with n observations for our problem is presented in Equation 3-5:

$$\boldsymbol{hr_i = b_0 + b_1 f_{i1} + \ ... + b_9 f_{i9} \qquad for\ \ i = 1, 2, ..., n} \qquad \textbf{3-5}$$

The goal is finding the best values for the regression coefficients (b_j) to fit the line to the observed data. The best fitting line is typically calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line. After building this model using observed data points, the unknown HR value could be estimated using observed features.

k-Nearest Neighbour (kNN) is a type of instance-based learning algorithm typically used for classification problems. It uses a simple distance measure to find the training instance closest to the current test instance and considers the same class as this training instance (Aha, Kibler, & Albert, 1991). However, if the target parameter (e.g., HR) has a numeric value, the kNN would be considered as a regression problem. In this case, the *k* neighbours nearest to the test instance are selected first, and then the average of their target values is assigned to the test instance. In our problem, suppose we have training data $(F_1, hr_1), ..., (F_n, hr_n)$, where $F_1, F_2, ..., F_n \in \mathbb{R}^p$, (p = number of features), $hr_1, hr_2, ..., hr_n \in \mathbb{R}$. If F^* is a test vector where $F^* \in \mathbb{R}^p$, we can predict hr^* by finding the k nearest neighbours of F^* and then computing the mean of the k nearest training values $(hr_{r1}, hr_{r2}\ , ..., hr_{rk})$ as shown in Equation 3-6.

$$\boldsymbol{hr}^{*} = \frac{\mathbf{1}}{\boldsymbol{k}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{k}} \boldsymbol{hr}_{\boldsymbol{r}_{\boldsymbol{i}}} \qquad \textbf{3-6}$$

The training and testing processes for both methods (Linear Regression and kNN) are performed separately with a 10-fold cross validation approach. The Waikato environment for knowledge analysis Weka (Witten & Frank, 2005) is used for executing these techniques.

The mean absolute error (MSE), root mean squared error (RMSE), and Pearson's correlation coefficient are calculated for the estimated HR and actual HR extracted from the reference ECG. Bland–Altman plots (Bland & Altman, 1986) are used for comparing proposed methods and actual HR values. The mean differences with 95% limits of agreement (± 1.96 SD) are also reported for each method. The Limits of Agreement (LoA) specify a range that most of the measurement errors lie within.

3.4 Classification of affective states

3.4.1 Feature selection

The methods described extract a large number of features. Some of them might not be related to affective states and some of them might provide some kind of redundancy. A feature selection technique needs to be applied on the extracted features before classification to remove unnecessary features. In addition, analysing the selected features could give us a better understanding about the relation between certain features and each affective state.

There are two broad categories for feature selection algorithms: wrappers and filters. Wrappers try to search throughout the features space and use a learning algorithm to build a model for evaluation. This approach is expensive computationally and might lead to an over fitting problem. On the other hand, filter methods evaluate the features according to a simple filtering measure instead of using a model-based evaluation. This approach is much faster than wrappers, especially when large data-bases are evaluated. The Correlation-based Feature Selection (CFS) (Hall, 2000) method is one filter algorithm that has been tested successfully on various applications and proposed for both discrete and continues features.

3.4.1.1 *Correlation-based Feature Selection (CFS)*

The CFS evaluates different possible subsets of features and ranks them based on its measure. The CFS measure works on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other" (Hall, 2000). This measure is calculated using following equation (Equation 3-7):

$$\mathbf{Merit_s} = \frac{\mathbf{k}\,\overline{\mathbf{r_{cf}}}}{\sqrt{\mathbf{k} + \mathbf{k}(\mathbf{k}-\mathbf{1})\ \overline{\mathbf{r_{ff}}}}} \qquad \mathbf{3\text{-}7}$$

where $\mathrm{Merit_s}$ is the measure of a subset feature S that contains k features; $\overline{\mathrm{r_{cf}}}$ specifies the average correlation between features; and class $\overline{\mathrm{r_{ff}}}$ indicates the average correlation between each pair of features. The correlation between the subset and the target class is measured in the numerator of that equation. The denominator could be a measure for showing the level of redundancy between the features. Equation 3-7 could evaluate different combinations of features and find efficient subsets of features for classification.

3.4.2 Classification

Various algorithms and methods have been proposed for affect classification (Calvo & D'Mello, 2010) and multiple studies have aimed to find the best algorithm for affect classification (Hussain et al., 2012b; Monkaresi, Calvo, et al., 2012). Some of them were successful in classifying affect using facial features but were not suitable for physiological features. To be more precise, introducing a single classifier that performs well for multiple modalities is difficult. In this thesis, we have two different modalities (facial expression and physiological signal) and utilizing a single classifier cannot produce an appropriate result. One solution is using combining classifiers. According to Mangai et al. (2010), there are two main reasons for combining classifiers. First, when there is a combination of features from different types and modalities, a specific classifier cannot perform well. In this occasion, combining classifiers can improve the performance of classification. The second reason is increasing the generalization power of the classification. Sometimes a specific classifier fails when it is faced with a new set of test data beyond the training data set. Our previous study (Hussain et al., 2012b) also suggested that combining classifiers could increase the accuracy of affect detection in multimodal systems.

Techniques that combine classifiers use a collection of individually trained classifiers to predict the classification of an instance by combining their individual predictions. There are two main strategies to combine these individual classifiers: *classifier fusion* (L. I. Kuncheva, 2004) and *classifier selection* (L I Kuncheva, 2002). In the classifier fusion strategy, all classifiers are provided the same input from the feature space. The result is calculated by combining the individual results predicted by each classifier. However, in the classifier selection strategy, each classifier is known as an expert in each specific domain and each specific part of the feature space is dedicated to each classifier. Another classifier is responsible for combining the result and makes the final decision.

We used the classifier fusion strategy for our problem because it is not clear which classifier is the expert in classifying affect using our feature space. Three techniques are proposed for fusing classifiers based on the type of classifiers outputs. If the outputs are crisp labels, the abstract level techniques, such as majority vote or weighted majority vote, are used. The "rank level" or "measurement level" techniques are used when the outputs of classifiers are a "subset of possible matches" or "probabilistic confidence measures", respectively (Mangai et al., 2010).

In this study the *vote classifier* with *the average probability rule* is used for combining the base classifiers. For the base classifiers, we selected the common types of classifiers widely used and showed reasonable performance in affect classification (Hussain et al., 2012b; Nguyen, Bass, Li, & Sethi, 2005). Support Vector machine (SVM), (kNN, and Decision Trees are considered as base classifiers. Figure 3-8 illustrates the overall process of combining classifier with an ensemble of three classifiers.

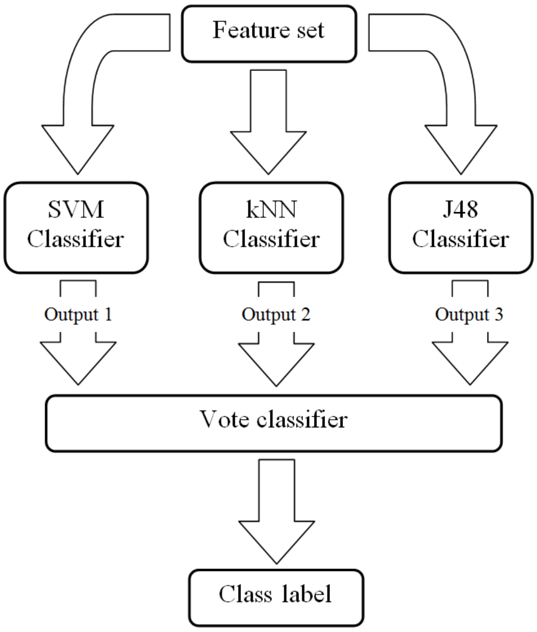


Figure 3-8: The classification process

3.4.2.1 *Support vector machines (SVMs)*

SVMs are powerful supervised learning algorithms used widely in different applications, including FER (Bartlett et al., 2005; Michel & El Kaliouby, 2003; Valstar, Mehu, Jiang, Pantic, & Scherer, 2012; Zhao & Pietikäinen, 2007). Vladimir Vapnik invented the original

SVM algorithm in 1995 (Cortes & Vapnik, 1995). The basic format of a SVM gets a set of input data and makes a model to distinguish two separate classes. This model predicts whether a new instance belongs to the target class or not. It can also be applied on multiclass separation problems.

The SVMs show high accuracies in a wide variety of applications and theoretically, they can deal well with the over-fitting problem. According to Michel and El Kaliouby (2003), the SVMs outperform neural networks in several applications. The main advantage of a SVM is the capability to learn from small datasets and generalize the model accurately. The size of datasets in the FER systems is typically small and SVMs are suitable in this application. The facial images also contain lots of noise due to head and body movements and illumination changes. The SVMs perform well in dealing with noisy data because of their generalization performance and the ability of separating samples, which are difficult to separate (Michel & El Kaliouby, 2003).

A SVM tries to construct a hyper-plane or a set of hyper-planes in a higher dimensional space as a decision boundary between two classes. The goal is finding an optimize hyper-plane by maximizing the margin (distance between the boundary and nearest instance). Figure 3-9 shows an example of a linear decision boundary. Sometimes the data are not spreadable using a linear equation. For these cases, the kernel tricks (Boser, Guyon, & Vapnik, 1992) can be applied on feature space to create a non-linear decision boundary and maximize the margin. The kernel trick is a way of transforming observations from a general set into an inner product space to make them linearly separable. Figure 3-10 shows a transformation of nonlinearly separable classes into a linearly separable classes using the kernel function φ. Some common kernel functions include Polynomial, Gaussian radial basis function, and Hyperbolic tangent.

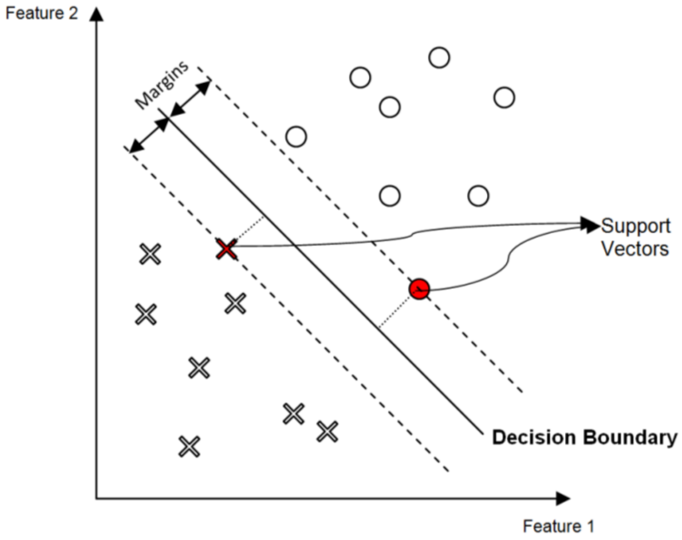


Figure 3-9 : An example of SVM classification of two classes (circles and crosses) for the separable case. The red instances are the support vectors.

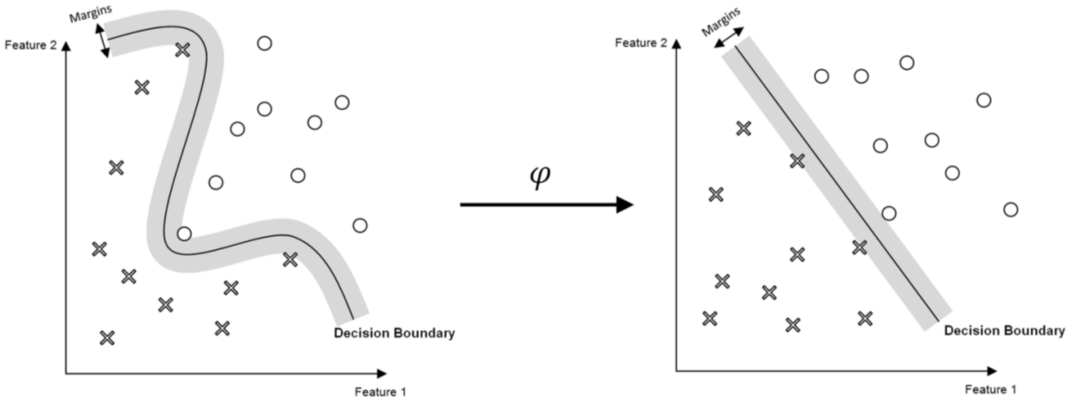


Figure 3-10: An example of mapping nonlinearly separable classes into linearly separable form. (Adapted from Polikar (2006))

In general, SVMs are considered a binary classifier; however, they can be applied on multiclass classification problems by reducing that into multiple binary problems. There are different reduction techniques, such as all-against-one or one-versus-one techniques. In all-against-one approach, n (n = number of classes) distinct classifiers need to be built for each class that can distinguish between the instances that belong to that specific class and the rest of labels. Each classifier gives a score to each instance and finally the label of that instance is detected by comparing the scores. The new instance belongs to the class with the highest score. However, in the one-versus-one approach, n × (n – 1) / 2 distinct classifiers for each pair of labels are built. To classify a new instance, each one-versus-one classifier is applied on the instance and the number of votes to each label will be computed. Finally, the label that received the maximum number of votes is assigned to that instance.

We use the Sequential Minimal Optimization (SMO) algorithm to train the SVMs implemented in the Weka package. This algorithm reduced the complexity and increased the speed of the training process by breaking very large quadratic programming (QP) optimization problems into a set of smaller QP problems that can be solved using analytical methods (Platt, 1998). A linear kernel function is used for the SVM and a one-versus-one approach is used in this implementation.

3.4.2.2 *k-Nearest neighbour (kNN)*

The kNN algorithm is a type of instance-based or lazy learning method to classify samples based on closest training examples. In lazy learning algorithms, the classification task is postponed until the new query is received, whilst in the eager learning, the method tries to make the training model before receiving new classification queries (Witten & Frank, 2005). The kNN algorithm is used widely in affect classification problems using different modalities, such as physiological signals (Wagner et al., 2005), facial expressions (Bourel, Chibelushi, & Low, 2002), and speech (Lee & Narayanan, 2005).

The kNN algorithm is very simple to implement, and it works well in basic classification problems. It can be applied on the feature space that is not linearly separable (Hand, Mannila, & Smyth, 2001). The main issue with this algorithm is that it needs a large memory space. In fact, it keeps all training instances in its memory to make its judgment about new examples.

This problem could be solved using advanced implementation of kNN, such as IB1 and IBk (Aha et al., 1991). The IBk implementation showed its robustness in dealing with noises and irrelevant attributes (Aha et al., 1991). We use this implementation of kNN in our proposed system, which is presented in the Weka classification library (Witten & Frank, 2005).

One of the main variables of a kNN algorithm that needs to be defined is the similarity function. This function defines the closeness measure of each training instance and new examples. The common similarity function is Euclidean distance, which can be used for continuous variables. To calculate the similarity between two discrete variables, other methods, such as the overlap metric or Hamming distance (Hamming, 1950), can be used. In the Euclidean method, the distance between two instances is calculated using the ordinary method for measuring distance between two points and is given by the Pythagorean formula. For example, in the n dimensional space, if $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$ are two points in Euclidean n-space, then the distance between p and q is given by:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad \textbf{3-8}$$

Another important parameter for each kNN implementation is the number of nearest neighbouring ks to explore. The label of each nearest neighbour is considered in a voting mechanism and finally the majority of nearest labels will select the label of the test sample. Figure 3-11 shows the impact of selecting two different values for the k parameter in a 2D space. The task is classifying the test sample represented by a white cross. When the kNN (k = 3) is applied to this example, the test sample is classified as the green circle class. If the kNN (k = 7) is applied to this example, the test sample is assigned to the red square class.

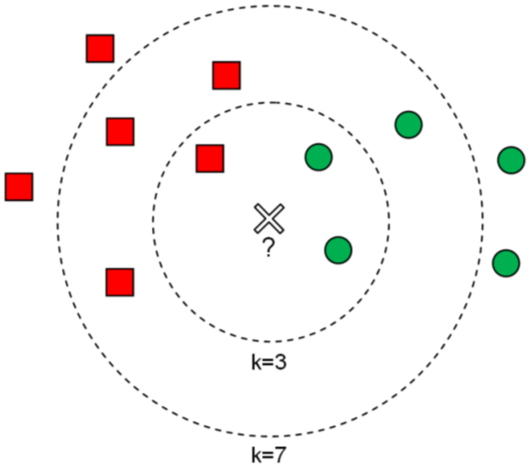


Figure 3-11 : An example of kNN classification using *k* = 3 and *k* = 7

The basic kNN classifier with the majority voting mechanism could be biased toward the more frequent classes, especially when the class distribution is skewed. It means the samples of the most frequent class are likely to appear among the *k* nearest neighbours. Using a weighted kNN could address this problem by considering the actual distances in the voting mechanism.

3.4.2.3 ***Decision Trees***

The Decision Tree algorithm is one of the most common methods in machine learning. It is generally used to support the decision making process based on the tree structure (Rokach & Maimon, 2008). It can be used for classification problem by considering each interior node as a condition on each attribute. The leaf nodes of the decision tree represent the label of each class. Compared with the black box models, such as neural networks, the tree structure provides a white box model that is simple to understand and can evaluate the impact of each attribute during classification process. Another advantage of decision trees is that they can perform well even with small number of data. Both numeric and discrete features can be evaluated using decision trees. An example of a simple decision tree is illustrated in Figure 3-12. In this example, the decision tree tries to classify the costumers of a company for advertising purposes. The internal nodes (circles) represent the conditional attributes and the leaf nodes (triangles) represent the class label. The classification process is started from the root of the tree. The navigation through the tree is continued based on conditions until a leaf is reached.

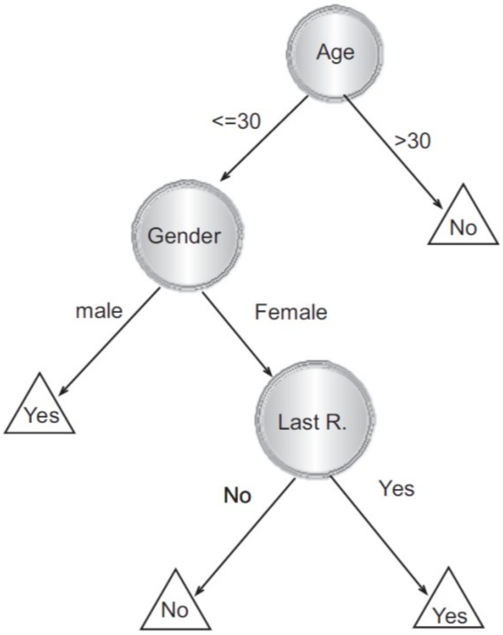


Figure 3-12: An example of a decision tree (Source : Rokach & Maimon (2008), p. 10)

Decision trees are applied successfully on different applications, including affect detection. For example D'Mello et al. (2007) achieved an accuracy of 77% for detecting frustration by applying decision trees on features. Nguyen et al. (2005) suggested that combining decision trees with SVM classifiers could improve the accuracy of emotion detection. In another study (Barreto, Zhai, & Adjouadi, 2007), the results showed a reasonable accuracy for decision trees to discriminate two affective states (stress vs. relax) using physiological signals.

One of the most common classification algorithms based on decision trees is called C4.5 developed by Ross Quinlan (1993). The C4.5 creates the decision tree based on training samples using information entropy metric (Borda, 2011). It starts with a set of training samples and tries to find the attribute that can split the training samples into the known class labels effectively. It uses the normalized information gain (difference in entropy) as a metric for selecting the most effective attribute in each iteration. This process is repeated on the smaller subsets to build the final decision tree. If all the samples in the sub-set belong to the same class, the iteration is stopped and a leaf with the same label is created. In this thesis, we utilize an open source Java implementation of the C4.5 algorithm in the Weka package (J48).

3.4.3 Performance metrics (F1-score, Kappa)

Several metrics are proposed for evaluating the performance of a classification algorithm. The most common metric is the *Accuracy* of a classifier that is measured by the proportion of correctly classified instances to the total number of test instances. This metric is useful but does not represent some valuable information, in particular when the class distribution is skewed (He & Garcia, 2009). For example, 90% of samples belong to class *A* and 10% of them belong to class *B*. In this case, if the classifier simply assign the same label *A* to all instances, it can achieve a reasonable accuracy (Accuracy = 90%). This example shows the weakness of the *Accuracy* metric in representing the performance of a classifier. To have a better understanding of the performance of a classifier other metrics, such as Precision, Recall, F-measure, and Kappa measure, were proposed.

3.4.3.1 ***Precision and recall***

Precision and recall are two well-known metrics in pattern recognition and information retrieval. To explain the precision and recall in the classification context, the confusion matrix needs to be defined first. The result of each classification task can be presented in a confusion matrix. The rows of the matrix represent the actual classes and the columns represent the predicted class labels using the classifier as shown in Table 3-3.

Table 3.3: The confusion matrix. Adopted from Witten and Frank (2005) p.162

		Predicted class	
		+	–
Actual class	+	TP (true positive) Correct result	FN (false negative) Missing result
	–	FP (false positive) Unexpected result	TN (true negative) Correct absence of result

According to Table 3-3, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are calculated by comparing the predicted results and the actual values. TP and TN represent correct classification whilst FP and FN represent a misclassification results. The precision is the number of correctly classified samples as belonging to the class (TP) divided by the total number of all samples classified as belonging to the class, no matter if they are labelled correctly or incorrectly as the positive class (TP + FP). Precision can represent the exactness of the classification task. However, recall can measure the completeness of the prediction by the ratio of the correct classified samples as the positive labels divided by the total number of the total number of samples that actually belong to the positive class. The actual number of samples belonging to the positive class includes the number of correctly classified positive samples (TP) and the number of misclassified samples as the negative class (FN). The precision and recall can be calculated using Equations 3-9 and 3-10:

$$Precision = \frac{TP}{TP + FP} \qquad \textbf{3-9}$$

$$Recall = \frac{TP}{TP + FN} \qquad \textbf{3-10}$$

If the precision value is equal to 1, it means that the classifier detected all positive samples correctly and did not classified any negative samples as the positive class. The value of 1 for the recall metric indicates that the classifier was 100% successful in classifying positive samples and did not misclassify any positive samples as the negative class. Both precision and recall metrics should be considered at the same time to have a reliable assessment of the classification performance.

The F1 score (F-measure) is another popular metric that takes these metrics into account in its measurement. In fact, the F1 score is the harmonic mean of precision and recall and is given by Equation 3-11:

$$F1\ score = 2 \cdot \frac{Precision\ \cdot Recall}{Precision + Recall} \qquad \textbf{3-11}$$

The F1 score ranges from 0 to 1. The value of 1 indicates the highest performance of the classifier. This metric can be a reliable alternative of the accuracy metric as it is less sensitive to skewed class distributions (Joshi, 2002). The F1 score also can be used for measuring the performance in multiclass detection problem. To do this, the F1 score for each individual class is calculated and the average of these values is considered as the final F1 score.

In this thesis, we evaluate the performance of the proposed system using F1 score besides other metrics because we anticipate that we might face with imbalanced class distribution in our experiments.

3.4.3.2 *Cohen's Kappa*

Another measure used commonly for measuring the performance of a classifier is Cohen's Kappa. It is used widely in affective computing applications (D'Mello & Graesser, 2010; M. E. Hoque, Kaliouby, & Picard, 2009; McDaniel et al., 2007; Tian et al., 2001; Valstar & Pantic, 2012). The Cohen's Kappa was introduced originally for measuring the inter-rater or inter-annotator agreement of those observing the same phenomenon (J. Cohen, 1960). Compared with other simple agreement measures (e.g., percent), the Cohen's Kappa is a robust measure because it is less sensitive to the agreement occurring by chance. The Cohen's Kappa is typically proposed for measuring the agreement between two annotators. Other metrics like the Fleiss's Kappa (Fleiss, 1971) is used when the number of annotators is more than two.

In the FER domain, researchers always use self-reports or ask experts to annotate an image or a video segment. This information is considered as the ground truth for the classification task. The Cohen's Kappa is a good measure to assess the level of agreement of annotators to build a reliable data-set. The Cohen's Kappa can measure the classification performance by measuring the level of agreement between the classifier prediction and the actual class labels. We will refer to the Cohen's Kappa by the term "Kappa measure" in the rest of this thesis. The Kappa measure is calculated using the following formula:

$$Kappa = \frac{\mathrm{Pr}_{(\mathrm{a})} - \mathrm{Pr}_{(\mathrm{e})}}{1 - \mathrm{Pr}_{(\mathrm{e})}} \tag{3-12}$$

where $\mathrm{Pr}_{(\mathrm{a})}$ refers to the relative agreement among raters, and $\mathrm{Pr}_{(\mathrm{e})}$ represents the probability of chance agreement. A Kappa value of 1 indicates the prefect agreement between two raters, whereas the value of 0 indicates that there is no agreement between raters. To have a better understanding in calculating the Kappa measure in the classification context, it is useful to explain it through the confusion matrix.

Table 3.4: Confusion matrix

		Predicted class		
		+	–	Total
Actual class	+	20	5	25
	–	10	15	25
	Total	30	20	50

Suppose that the results of a classification task are reported in Table 3-4. According to this table, 20 positive and 15 negative samples (out of 50 samples) are classified correctly, so the total agreement probability is:

$$\mathrm{Pr}_{(a)} = \frac{20 + 15}{50} = 0.7$$

To compute the probability of chance agreement between predicted and actual class, Equation 3-13 needs to be calculated:

$$\mathrm{Pr}_{(e)} = \frac{\mathrm{N_{ap}} \times \mathrm{N_{pp}}}{\mathrm{N}} + \frac{\mathrm{N_{an}} \times \mathrm{N_{pn}}}{\mathrm{N}} \qquad \textbf{3-13}$$

where N is the total number of samples; $\mathrm{N_{ap}}$ is the number of actual positive samples‘ $\mathrm{N_{pp}}$ is the number of predicted positive samples; $\mathrm{N_{an}}$ is the number of actual negative samples; and $\mathrm{N_{pn}}$ is the number of predicted negative samples. Hence, the $\mathrm{Pr}_{(e)}$ for the mentioned example is:

$$\mathrm{Pr}_{(e)} = \frac{25 \times 30}{50} + \frac{25 \times 20}{50} = 0.5$$

Having $\mathrm{Pr}_{(a)}$ and $\mathrm{Pr}_{(e)}$, the Kappa measure is given by:

$$Kappa = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

The Kappa measure of 0.4 represents a fair agreement between actual and predicted class labels. This metric is also useful in assessing the performance of the classifier when the class distribution is skewed. In general, a Kappa measure between 0.4 and 0.6 is considered as a fair accuracy, and the value greater than 0.75 is considered as an excellent agreement (Robson, 2011). D’Mello et al. (2007) achieved the maximum kappa measure of 0.54 for discriminating flow and boredom using the user’s posture pattern during interaction with an intelligent tutoring system. Alzoubi et al. (2012) also proposed a model to detect eight affective states (boredom, confusion, curiosity, delight, flow/engagement, surprise, and neutral) using physiological signals and this obtained the mean Kappa measure of 0.25.

In this thesis, the Kappa measure is also reported to evaluate the performance of the proposed system and reliability of class labels.

3.4.4 Validation

In this section, we explain the techniques for evaluating the generalization power of the prediction model. There are several possible ways to select training and testing data and evaluate the performance of the prediction model. Two main approaches are described here: k-fold cross validation and leave-one-out cross validation approach.

3.4.4.1 *k-fold cross validation*

In k-fold cross validation, the dataset is randomly divided into *k* folds with equal sizes. In each iteration, one fold is considered as the test set and the model is trained by the rest of *k–1* folds. This process is repeated *k* times for each fold separately, and the performance metric of the system is stored for all evaluations. The final metric can be calculated by averaging over the *k* results (Witten & Frank, 2005). The value of *k* could be varied in different application. Ten-fold cross validation is the most common technique for evaluating the performance of a classifier. Using various combinations of training and testing samples could represent the generalization power of the system. This approach could prevent the biased evaluation that might occur due to using a bad configuration of training and testing sets (e.g., overlapping in training and testing datasets).

3.4.4.2 *Leave-one-out cross validation*

This approach could be considered as a specific type of *k*-fold cross validation. In this case, the number of folds (*k*) is equal to the number samples (Polikar, 2006). In some classification problems, the dataset is grouped by certain parameters. For example, in user-independent analysis, the dataset could be divided into the portions, including individual data. The LOO cross-validation can be utilized on the user level. To be more precise, if the dataset is a combination of N samples from M users, we can leave the portion of one user as the test set and use the rest of the dataset as the training set. Similar to the k-fold cross validation approach, this process is also repeated M times. This technique is also called as "Leave-one-subject-out" cross validation. Due to intentional separation of the dataset, we can estimate the performance of our system in the real world when it is faced with unknown users.

In this report, we use the 10-fold cross validation technique to validate the classification accuracies in user-dependant analysis. For the user-independent analysis, we use the leave-one-subject-out technique to have a more reliable estimation of the performance of our system in predicting affective states.

3.5 Corpora

In affective computing research, existing databases are mostly focused on basic emotions (Afzal & Robinson, 2009). Most of these databases were recorded in a controlled environment and the emotional scenes were played by actors (M. E. Hoque et al., 2012). Several databases have been developed for detecting affective states in HCI applications. However D'Mello and Calvo (2013) showed that non basic emotions are more important but less investigated in the HCI applications. They have provided evidence that four non-basic affective states (engagement, boredom, confusion, and frustration) are the most common in the interaction between human and computer. However, eliciting and analysing naturalistic data are complex and expensive. To evaluate our proposed system to detect naturalistic complex emotions, we need to create a new dataset.

In this section, we explain the method for collecting *naturalistic* expressions of *non-basic emotions* during human and computer interactions. New less-intrusive methods and sensors to record user's behaviour and expressions could improve the accuracy of affect detection. Hence, in this study, the Microsoft Kinect sensor is used to record depth, video, and audio simultaneously.

3.5.1 Procedure

The experiment consists of two main parts. In the first part, the participants were required to write about a given topic using a computer-based system. In the second part, they were asked to watch a certain number of standard emotional images from the International Affective Picture System (IAPS) (P. Lang & Bradley, 1997) collection and report their affect. The second part takes about 30 minutes. The detailed procedure for each part is conducted as described in following subsections. Table 3-5 provides the summary of the data acquisition protocol.

Table 3.5: Overview of our data acquisition protocol

	Part	Duration	Coders	Segmentation	Coding time	Annotation scheme	Affect content
1	**Writing**	60 min	Self-report	Interval-based	Concurrent	Categorical	Engagement
				Event-based	Retrospective	Categorical	Emotion list *
				Event-based	Retrospective	Dimensional	Valence/ Arousal
2	**IAPS**	30 min	Normative	Interval-based	Pre-defined	Dimensional	Valence/ Arousal
			Self-report	Interval-based	Concurrent	Categorical	Emotion list*
				Interval-based	Concurrent	Dimensional	Valence/ Arousal

* Emotion list: Calm, Relaxed, Bored, Annoyed, Glad, Content, Delighted, Excited, Depressed, Gloomy, Afraid, Angry and Other.

3.5.1.1 ***Part 1: Writing session (60 minutes)***

In our writing studies, participants were asked to write a travel piece (journalistic genre) about a location that they visited recently. The task required some research but not much prior knowledge, and feedback could be provided that triggered higher arousal emotions. The writing session is based on a *draft-review-final* activity, and after receiving the topic:

1) They write a draft and submit it (30 minutes).

2) They wait for 10 minutes to receive feedback. They are asked to stay seated while feedback is being processed but they are free either to work on other manuscripts or browse the internet.

3) They receive human and automated feedback. The feedback contains suggestions on how to improve the quality of the writing.

4) They have an additional 20 minutes to revise their manuscript according to the feedback and submit the final version.

During the writing session and every two minutes, the system produces an *auditory probe* to notify the subject to report its level of engagement verbally. A list of examples is given to the participant before the session and they just need to say the corresponding affective state. The voice is also recorded and analyzed later. This method is one of the most simple and non-intrusive methods for concurrent self-reporting, which provides more accurate and valuable annotation compared with interrupting the recording session by traditional questionnaires (Afzal & Robinson, 2009).

We utilize a web-based system called *Tracer* (M. Liu, Calvo, & Pardo, 2013) that manages the writing activities and records the writing process, saving versions of the document every 20 seconds. This text-in-time data allows us to mine some of the cognitive processes associated with writing from the patterns of text additions, edits, deletions, and pauses. Distracting noises (e.g., someone coming in the office, telephone ringing) are produced and annotated during the session to simulate more a naturalistic environment.

3.5.1.2 *Part 2: IAPS images (30 min)*

A total number of 60 images for 10 seconds each one is presented, followed by 10 seconds pauses between the images for annotation. The images are selected from the IAPS database. It provides a set of emotional stimuli to trigger emotion and attention, and it is widely used in affective computing research.

Each picture in this database was rated by approximately 100 participants in terms of three dimensions: valence, arousal, and dominant (P. J. Lang et al., 2008). A 1–9 rating scale was used for each dimension. The ratings were averaged and provided with the pictures. The corpus also provides gender specific ratings. Different patterns in the ratings are observed for each gender. For example the “EroticFemale” picture (image number 4235) was rated as a normal picture by female participants (v = 3.67, a = 3.97), whereas it was rated as a highly emotional pictures by male participants (v = 7.29, a = 6.73). It shows the importance of considering *gender* for designing emotional stimulus in affective computing research. We used these ratings to organize two gender specific groups for our experiment. Each gender specific group was divided into four categories based on arousal and valance normative rates. Table 3-6 represents the thresholds for selecting images in each category. Fifteen images were selected for each category based on the criteria presented in Table 3-6. More specific information about the selected pictures in each category is provided in Appendix A.

Table 3.6: Used criteria for selecting images from the IAPS database

	Categories	Male		Female	
		Valence	**Arousal**	**Valence**	**Arousal**
	Available range in IAPS	[1.5, 8.39]	[1.55, 7.8]	[1.15, 8.74]	[1.87, 7.77]

1	**Low Valence - Low Arousal**	< 4.0	< 4.0	< 4.0	< 4.1
2	**Low Valence - High Arousal**	< 4.0	> 5.71	< 4.0	> 5.81
3	**High Valence - Low Arousal**	> 6.1	< 4.0	> 6.21	< 4.1
4	**High Valence - High Arousal**	> 6.1	> 5.71	> 6.21	> 5.81

3.5.2 Segmentation

After the experiment, various processes are done on the gathered data. First, the video is divided to certain meaningful segments for annotation. *Interval-based* and *event-based* segmentation are the two common types of video segmentation. Each method has its strengths and weaknesses. Choosing the method of video segmentation depends on the nature of the data. The behaviour of the user is unpredictable during writing in the first part of the experiment. In these conditions, researchers (Afzal & Robinson, 2009; Mahmoud et al., 2011) have suggested that event-based segmentation provides more useful annotations. Accordingly, the event-based segmentation is considered for segmenting recorded videos in writing sessions. In this case, each segment shows a single event, such as a change in facial expression, head, and body posture movement.

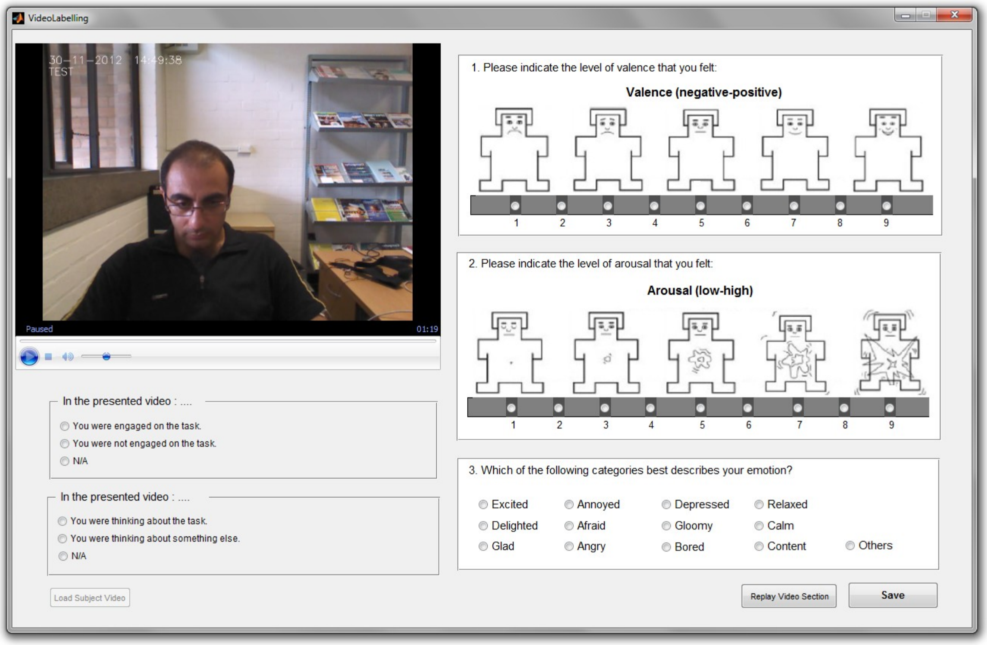
However, interval-based segmentation is suited for the second part of the experiment (IAPS session) because the time and the length of each interval could be defined corresponding to the time of each presented image. So, 60 ten-second video segments could be produced from the second part of the experiment (IAPS images).

3.5.3 Annotation

3.5.3.1 *Part 1*

The participant is asked to annotate their recorded video when the segmented videos are ready. These annotations are elicited after the writing session as a *retrospective* report. They are free to play each segment as many times as needed to form a better judgment. The questionnaire asks the following questions regards to engagement: *Were you engaged in the* *task or not?* and *Were you thinking about the topic or not?* These questions are designed based on the experience sampling study by Killingsworth and Gilbert (2010).

The second group of questions related to the participant emotion. A 1–9 scale of valence and arousal was collected using the Self-Assessment Manikin (SAM) method (P. Lang & Bradley, 1997). A set of affect categories were provided for the participant: Calm, Relaxed, Bored, Annoyed, Glad, Content, Delighted, Excited, Depressed, Gloomy, Afraid, Angry, or "Other". In addition, the emotions that had been said aloud during the session by the subjects were extracted from the recordings. Figure 3-13 shows a screenshot of the retrospective self-report questionnaire.



In the presented video : ....

○ You were engaged on the task.
○ You were not engaged on the task.
○ N/A

In the presented video : ....

○ You were thinking about the task.
○ You were thinking about something else.
○ N/A

1. Please indicate the level of valence that you felt:

Valence (negative-positive)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9

2. Please indicate the level of arousal that you felt:

Arousal (low-high)

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9

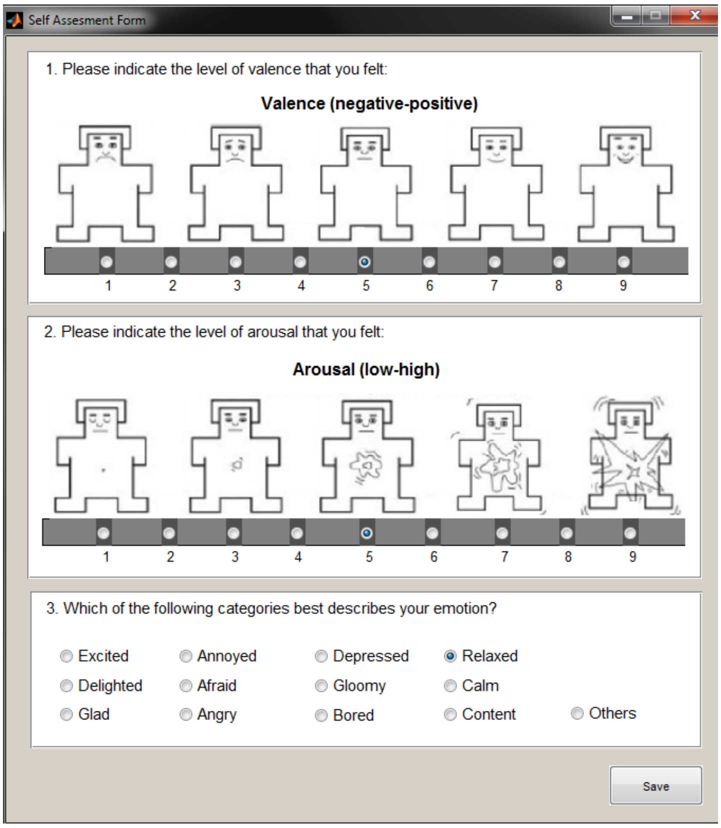
3. Which of the following categories best describes your emotion?

○ Excited ○ Annoyed ○ Depressed ○ Relaxed
○ Delighted ○ Afraid ○ Gloomy ○ Calm
○ Glad ○ Angry ○ Bored ○ Content ○ Others

3-13: A screenshot of the retrospective labelling questionnaire

3.5.3.2 *Part 2*

For the second part of the experiment, two types of annotation were provided for each video segment: *normative-rating* and *self-report*. The normative-rating was extracted from the IAPS dataset and this indicated the emotional level of each presented image in terms of valence and arousal, regardless of the user's reaction. In addition, for this part of the experiment, the self-reports were collected concurrently during the experiment. After viewing each image, the participants were asked to fill-out a questionnaire to indicate their degree of valence and arousal. The SAM protocol (P. Lang & Bradley, 1997) was used for this reason. They were also asked to report their affective states by choosing an affect category among provided categories. These annotations were used as the ground truth for evaluating our proposed system. Figure 3-14 presents a snapshot of the annotation page.



1. Please indicate the level of valence that you felt:

Valence (negative-positive)

○ 1 ○ 2 ○ 3 ○ 4 ● 5 ○ 6 ○ 7 ○ 8 ○ 9

2. Please indicate the level of arousal that you felt:

Arousal (low-high)

○ 1 ○ 2 ○ 3 ○ 4 ● 5 ○ 6 ○ 7 ○ 8 ○ 9

3. Which of the following categories best describes your emotion?

○ Excited ○ Annoyed ○ Depressed ● Relaxed
○ Delighted ○ Afraid ○ Gloomy ○ Calm
○ Glad ○ Angry ○ Bored ○ Content ○ Others

Figure 3-14: An example of the concurrent self-report form

Chapter 4. Study 1: Evaluating the video-based method for HR estimation

Summary

In this chapter, the proposed method for contact-less measuring of Heart Rate (HR) is evaluated under three different conditions. In the first experiment, we validated our implementation when the user was at rest. Then we evaluated and compared Poh et al.'s (2010) with our proposed method under two new conditions; naturalistic human computer interaction and exercising scenarios that contain more user motion, bigger dynamic range, and different lighting conditions. We analyzed the limitations of Poh et al.'s method and provided our own improvements using Machine Learning techniques.

4.1 Introduction

The recent work of Poh et al. (2010) for producing heart rate signals from a webcam has opened the opportunity to numerous new applications. This technique uses video of the face and Independent Component Analysis (ICA) (Comon, 1994) to detect core physiological signals such as heart rate and respiration. This technology can be used to help monitor the health parameters of both individuals with cardiovascular disease and healthy individuals in nonlaboratory conditions.

Poh et al. (2010) improved previous attempts (Takano & Ohta, 2007; Verkruysse et al., 2008), to develop a low-cost accurate video-based method for contact free HR measurement. The algorithm is based on a Blind Source Separation (BSS) technique that recovers unobserved signals or sources from a set of observed mixtures with no prior information about the mixing process. There are several techniques for BSS, including ICA (Comon, 1994) used by Poh et al. (2010). ICA is a technique for uncovering the independent source signals from a set of observations that are composed of linear mixtures of the underlying sources. The source signal in this study is the Blood Volume Pulse (BVP) that propagates throughout the body. High degrees of agreement (correlation coefficient r = 0.99) were achieved for measures of HR using the proposed method and HR extracted from the BVP sensor.

In this chapter we evaluated our proposed method for remote measuring heart rate described in Chapter 3 (See section 3-3-3) in different conditions and compared the results achieved by our method with the method proposed by Poh et al. (2010). Our evaluation consists of three experiments. First the implementation of Poh et al. (2010) method was replicated using a new dataset. In the second experiment, the accuracy of this method was evaluated on a long naturalistic HCI. Thirdly, the robustness of the method was also assessed under new conditions which included motion artifacts and a wide-range of HR changes.

For evaluation, the mean absolute error (MSE), root mean squared error (RMSE) and Pearson's correlation coefficient are calculated for the estimated HR and actual HR extracted from the reference ECG. Bland-Altman plots (Bland & Altman 1986) are used for comparing proposed methods and actual HR values. The mean differences with 95% limits of agreement (± 1.96 SD) are also reported for each method. The Limits of agreement (LoA) specify a range that most of the measurement errors lie within it.

4.2 Studies, participants and recording methods

4.2.1 First Experiment–Recording at rest

Ten volunteer participants (mean age=26.7 years, 8 males, 2 females, 80% Caucasian and 20% Asian) from The University of Sydney participated in the first and second experiment. All participants signed an informed consent form prior to data collection. This experiment was approved by the University of Sydney's Human Ethics Research Committee prior to data collection.

All participants were seated in front of the same computer running Windows XP in a normal indoor environment. Video recording was carried out using an ordinary webcam (Logitech Webcam Pro 9000) mounted on the screen. All videos were recorded in colour (24-bit RGB with 3 channels, 8 bits/channel) at 30 frames per seconds (fps) with pixel resolution of 640 × 480 pixels and saved in AVI format. In order to record physiological changes, electrocardiogram (ECG), respiration, and galvanic skin response (GSR) sensors were placed on participants’ bodies. BIOPAC MP150 system with AcqKnowledge (v. 3.8.2) software was used to acquire the physiological signals at a sampling rate of 250 Hz. Two electrodes were placed on the wrists for collecting ECG. The ECG-100C amplifier was used for ECG recording. GSR was recorded from the index and middle finger of the left hand and a respiration band was strapped around the chest to collect respiration rate. GSR and respiration rate data were not used in this experiment. The participants were asked to keep their movement to a minimum for one minute to record the baseline physiological signals. This part was used in the first experiment. The experiment was conducted in a normal room with normal artificial fluorescent ceiling light in combination with a varying amount of sunlight coming through windows from the left side of the participant.

4.2.2 Second Experiment–Naturalistic HCI

The second experiment was conducted using the same materials used in the first experiment. Immediately after completing the first experiment participants were asked to create their personal websites in Google sites. They were allowed to explore the Internet and use external resources for completing the tasks. During the interaction, video and ECG signals were recorded simultaneously. Each interaction lasted about 30 minutes (±10 min).

4.2.3 Third Experiment–Indoor Cycling

This experiment was conducted in an indoor gym environment with the participation of a female participant. The only illumination source was the ambient artificial fluorescent light. The same camera was mounted in front of a cycling machine for recording the participant's upper body. Two electrodes were placed on the participant's wrists and the earth electrode was placed on her left arm to record ECG. The participant's physiological signals were recorded while the participant was cycling (Figure 4-1).

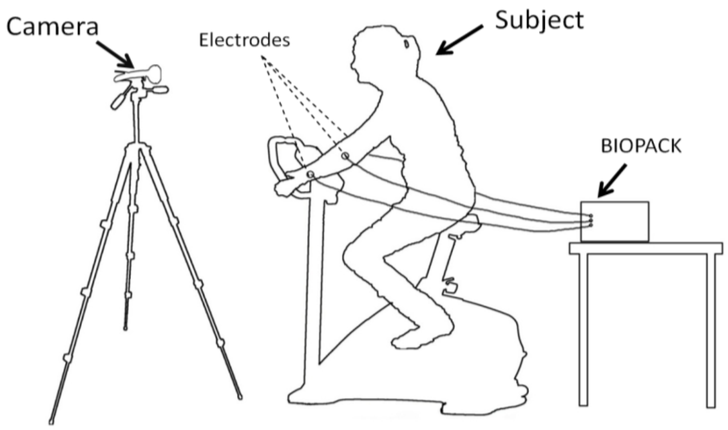


Figure 4-1: Experiment setup for the third experiment.

The experiment consisted of seven levels. In the first and last parts, the participant was seated on the cycling machine at rest position (no cycling) for two minutes. The other parts of the experiment were captured while the participant was cycling at different resistance levels (as shown in Table 4-1). The resistance of the cycling machine was gradually increased up to level 4 and returned to the easiest level. The participant was cycling for two minutes in the "Easy" levels and for three minutes in "Normal" and "Hard" cycling levels.

Table 4.1: Levels of activity in the indoor cycling experiment

Part	Level	Duration
1	Rest	2 min
2	Easy cycling	2 min
3	Normal cycling	3 min
4	Hard cycling	3 min
5	Normal cycling	3 min
6	Easy cycling	2 min
7	Rest	2 min

4.3 HR estimation results

In this section, we have reported the HR estimation results achieved after utilizing our proposed method which was described in the Section 3-3-3.

4.3.1 First experiment: validation

This study aimed to replicate the results of Poh et al. (2010) with our own implementation of the algorithm, and our own data. The first one minute of each participant's video recording in which their movements were kept to a minimum were analyzed at this stage. Thirty one estimations were performed for each participant. The first 30-second window was considered to estimate the HR at t=30. The standard deviations of the x- and y-coordinates of the face tracker of all participants were 4.56 and 2.57 pixels respectively. The average actual HR of 78.80 bpm was extracted from ECG signals among all participants. The standard deviation of extracted actual HR was 8.84 bpm (Table 4-2).

Table 4.2: The first experiment description: validation of the implemented method for estimating heart rate

	Poh et al.'s Study	First Experiment
Parameters		
Window size (seconds)	30	30
Experiment length (seconds)	60	60
# of participants	12 (2 F, 10 M)	10 (2 F, 8 M)
Recording rate (fps)	15	30
# of frames	10800	18000
# of measurements	372	310
SD of x- coordination (pixels)	-	4.56
SD of y- coordination (pixels)	-	2.57
Range of x- movement (pixels)	-	23
Range of y- movement (pixels)	-	18
Face Detection Acc	FN=0%, FP=0%	FN=0%,FP=0%
Mean actual HR	-	78.80
SD of actual HR	-	8.84

F=Females, M=Males, FN = false negative, FP = false positive.

Table 4-3 presents a comparison of our implemented algorithms applied in this experiment and Poh et al.'s experiment results. It should be mentioned that after ICA analysis, the third component produced the best results among all three components. The MPA method produced a similar performance. The results achieved from the third component are presented in Table 4-3. A correlation coefficient of 0.99 between estimated HR and the actual HR was achieved in this study. The RMSE of measurement at rest in (Poh et al., 2010) was 2.29, while in this study the RMSE was reduced to 1.69. Furthermore, using this experimental data, the mean bias was 0.86 bpm and the LoA span was 5.7 bpm, slightly better than Poh et al.’s (Poh et al., 2010) results in the sitting still experiment (mean bias = 0.05 bpm, LoA span = 8.99 bpm). Our results also showed reasonable accuracy compared to a new method for HR measurements at rest using Fiber Bragg grating-based sensor (Dziuda, Skibniewski, Krej, & Baran, 2013) (mean bias = - 0.01 bpm, LoA span = 3.64 bpm). Overall, the results showed that the accuracy of our implementation and this new dataset was comparable with Poh et al.'s report.

Table 4.3: Comparison of the results published by Poh et al. and our study using the same algorithm

	Poh et al.'s Study	**First Experiment**
Method	ICA	ICA
Selected component	2nd	3rd
Mean bias (bpm)	-0.05	0.86
SD of bias (bpm)	2.29	1.46
Upper limit (bpm)	4.44	3.71
Lower limit (bpm)	-4.55	-1.99
RMSE	2.29	1.69
Corr. coefficient	0.98***	0.99***

***: ($p<0.001$);

4.3.2 Second experiment: Evaluate the algorithm in HCI

The goal of this experiment was to evaluate the robustness and reliability of the video based, non-contact HR measurement in a naturalistic HCI scenario. For this reason, participants were free to move and no advice was provided regarding positions. The interaction with the computer was immediately started after one minute of a no-movement situation which was also used in the first study. We considered the whole session as one dataset. On average, each session lasted about 30 minutes but in order to compare across the same length for all participants the first 20 minutes of each session were analyzed here. This analysis has been done based on 30-second moving window. Table 4-4 provides more information about this study.

The face tracking algorithm was robust enough in detecting the face during head movement. However, sometimes because of the large degree of head tilting and turning (more than 45 degrees) or the occlusion of the face, the algorithm could not find the frontal face in the recorded images. In order to have a fair evaluation of the ICA analysis, the time windows which included false negative image frames were ignored for further analysis. Thus, among 11710 potential 30-second windows, only 9329 were selected for ICA analysis. The average range of the x- and y-coordinates of the detected faces was 64.00 and 35.5 pixels respectively.

Table 4.4: Second experiment description: heart rate estimation during human-computer interaction

	Second Experiment
Parameters	
Window size (seconds)	30
Experiment length (seconds)	1200
# of participants	10 (2 female, 8 male)
Recording rate (fps)	30
# of frames	360000
# of measurements	9329 (of 11710)
SD of x- coordination (pixels)	12.16
SD of y- coordination (pixels)	7.24
Range of x- movement	64.00
Range of y- movement	35.50
Face Detection Acc	FN=0.04%, FP=0%
Minimum actual HR	61.27
Maximum actual HR	111.79
Mean actual HR	80.55
SD of actual HR	9.52

4.3.2.1 *User-dependent analysis*

In this part, we have evaluated different components to estimate HR for each participant. The third component achieved the best results among eight out of 10 participants by comparing RMSEs. The MPA method improved the HR estimation for the remained two participants. The best results obtained by ICA analysis for each participant are presented in Table 4-5. In addition, the results achieved by applying kNN and linear regression are also presented in this table. The figures show the significant improvement of ICA with the kNN method over the former method (only ICA) for all participants. The regression did not improve the performance very much. These results imply the feasibility of building personal models for HR estimation during HCI in naturalistic scenarios.

4.3.2.2 *Combined-participant analysis*

Data from individual participants were combined to yield one large dataset. This dataset was used for training and testing the models with the 10-fold cross validation approach. On the other hand the HR was also estimated by the ICA methods. Here, the third component also provided the best performance among all three components. The descriptive statistics of HR estimation through different approaches are presented in Table 4-6.

Table 4.5: Results for hr estimation using the ICA method and the improved method by ml techniques (Participant-dependent analysis)

ID	Method	R	MAE	RMSE
1	ICA (3rd Comp.)	-0.38	8.67	21.16
	ICA + kNN	0.85	0.63	1.20
	ICA + Reg.	0.44	1.49	1.92
2	ICA (3rd Comp.)	0.10	2.79	4.34
	ICA + kNN	0.81	0.63	1.38
	ICA + Reg.	0.16	1.68	2.20
3	ICA (3rd Comp.)	0.29	46.83	59.13
	ICA + kNN	0.92	0.44	1.09
	ICA + Reg.	0.34	1.77	2.54
4	ICA (3rd Comp.)	0.41	45.34	57.39
	ICA + kNN	0.93	0.43	0.92
	ICA + Reg.	0.48	1.39	2.12
5	ICA (MPA)	0.18	33.92	41.98
	ICA + kNN	0.90	0.69	1.62
	ICA + Reg.	0.47	2.48	3.20
6	ICA (3rd Comp.)	0.29	3.96	8.20
	ICA + kNN	0.92	0.48	0.95
	ICA + Reg.	0.49	1.54	2.06
7	ICA (3rd Comp.)	-0.34	32.41	44.31
	ICA + kNN	0.83	0.72	1.19
	ICA + Reg.	0.50	1.37	1.77
8	ICA (3rd Comp.)	-0.02*	14.27	28.35
	ICA + kNN	0.86	0.58	1.18
	ICA + Reg.	0.46	1.51	1.98
9	ICA (MPA)	0.15	56.85	63.62
	ICA + kNN	0.81	0.84	1.75
	ICA + Reg.	0.18	2.27	2.81
10	ICA (3rd Comp.)	0.25	16.92	28.67
	ICA + kNN	0.91	1.38	3.10
	ICA + Reg.	0.49	5.15	6.37

*: $p<0.05$; $p<0.001$ for the rest; kNN(k=1); Reg.: Linear Regression; r: Pearson's correlation coefficient; MAE: Mean Absolute Error; RMSE: Root Mean Squared Error

Table 4.6: The descriptive statistics of hr estimation using the ICA method and improved method by ML techniques (combined-participants analysis)

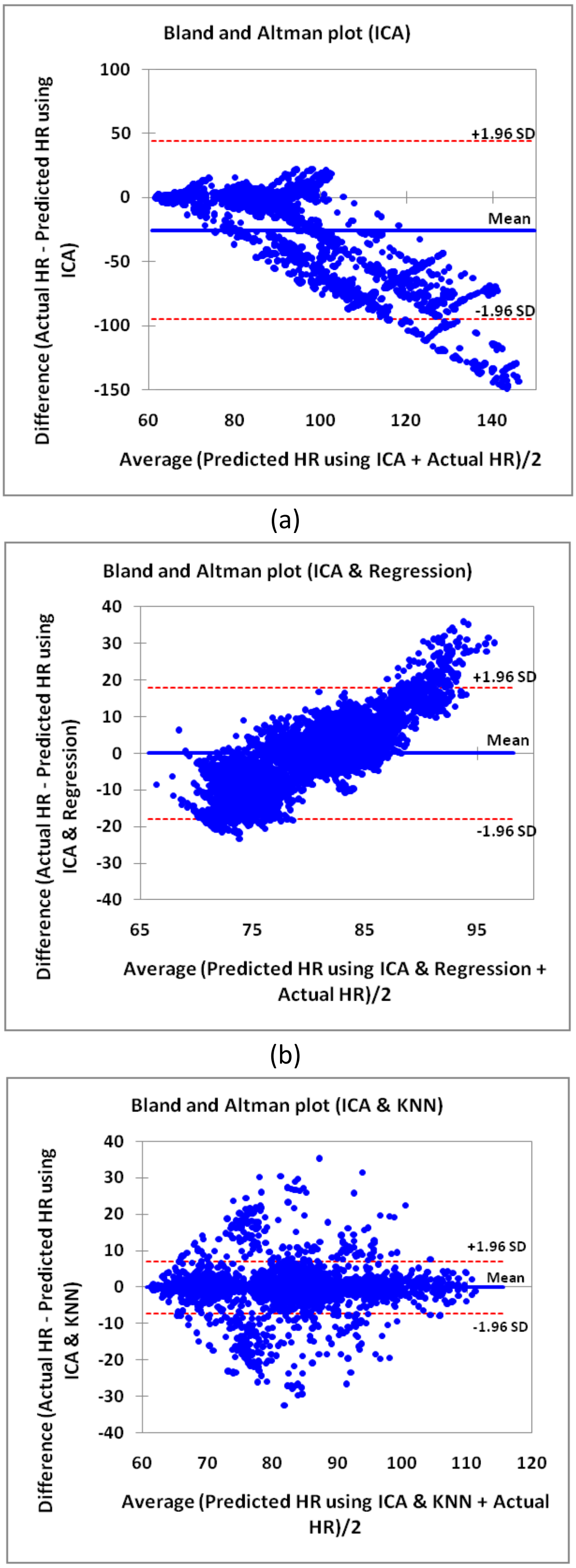
	Second Experiment		
Method	**ICA [1]**	**ICA + kNN**	**ICA + Regression**
Selected component	3rd	-	-
Mean bias (bpm)	-25.38	0.06	0.00
SD of bias (bpm)	35.65	3.65	9.10
RMSE	43.76	3.64	7.31
Corr. Coefficient	-0.10***	0.93***	0.29***
Mean absolute error	27.17	1.32	9.10
Relative absolute error	3.43	0.16	0.92
Root relative squared error	21.14	0.38	0.95

***: (p<0.001);

The large values for the mean bias -25.38 bpm (SD=35.65 bpm) would also indicate the weakness of the ICA method in detecting HR in the natural HCI environments. Applying ML techniques and the ICA method significantly reduced the absolute error of the HR estimation. The mean absolute error reduced from 27.17 bpm to 1.32 bpm and 9.10 bpm by applying kNN (k=1) and linear regression techniques respectively. The ICA and kNN technique also obtained a strong correlation with the reference sensor by reaching the Pearson's correlation coefficient of 0.93 (p<0.001).

The results of the Bland-Altman analysis for the three methods of HR prediction are presented in Figure 4-2. The LoA range was very wide for the ICA and for ICA and regression predictions. In contrast, the ICA and kNN method achieved the best accuracy with the LoA from -7.09 to 7.21 bpm (Figure 4-2(c)). In Figure 4-2(c), the difference values have more variations for mid-range average values (around 80 to 90) and smaller variations for small and large average values. One possibility is that there are more samples in the mid-range. That is, there are more people in this range and fewer people at two extremes. When there are more people, you would naturally have a wider range of difference values. However, there were no systematic or proportional errors observed for the ICA and kNN method as shown in Figure 4-2(c).

According to Figure 4-2(a), even though the actual HR range was between 61.27 and 111.79 bpm the ICA method predicted very large values most of the time. This might be because other strong frequencies appeared in the 3rd independent component. Figure 4-2(b) shows a case of proportional error for the ICA and regression method. It suggests that the relation between the nine features and actual HR might not be linear.



(a)

(b)

(c)

Figure 4-2: Bland-Altman Plots analysing the agreement between measured actual HR and measured HR using (a) ICA, (b) ICA with Regression and (c) ICA with kNN in the second experiment

4.3.3 Third experiment- Indoor exercising

In this study, the participant was asked to keep her movement to a minimum during the experiment. Obviously, some unwanted movement during the cycling was acceptable. The false negative ratio was 0.0.

We explored here the robustness of the proposed algorithm for a wide dynamic range of HR. In the second study, the average of the range and standard deviation of actual HRs among all participants were 16.50 bpm and 3.00 bpm respectively. In this study the HR of the participant started from 75 bpm, reached 130 bpm in the middle of the experiment and went down to 85 bpm. Figure 4-3 presents the actual HR changes during the experiment extracted from the ECG signals. Seven parts of the experiment are shown in the corresponding columns in Figure 4-3. The range and standard deviation of actual HR in the first, second and sixth parts were smaller than in the other parts.

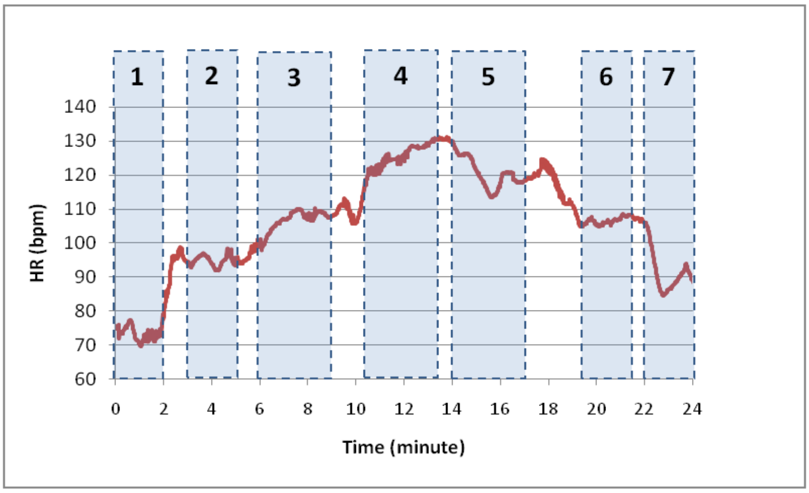


Figure 4-3: Actual HR changes during the third experiment (Indoor exercising) extracted from ECG.

The results of applying the proposed algorithm on these parts are presented in Table 4-7. The strategy for component selection in each part was to choose the component which contains the spectrum that minimizes the RMSE value. Consequently, the 3rd component was selected in part 4, the 2nd component was selected in part 5 and the 1st component was selected in the other parts. Our evaluations also stated that the HR spectrum could not be estimated accurately by considering the spectrum containing the highest peak among all three components.

As shown in Table 4-7, the ICA method achieves acceptable results in the first and second parts with RMSE values of 5.30 and 4.32. In these two parts the range of actual HR changes was smaller than other parts except part 6 which did not achieve a good RMSE. On the other hand, the algorithm showed the worst result in part 4 which contained a gradually increased HR.

Table 4.7: Summary of the third experiment results: HR estimation during indoor exercising using the ICA method

Parts	1	2	3	4	5	6	7
Moving window size	30 sec						
Recording rate	15 fps						
Length (sec)	120	120	180	180	180	120	120
# of measurements	91	91	151	151	151	91	91
SD of x- coordination	4.92	5.06	6.54	9.97	8.65	8.06	4.05
SD of y- coordination	5.14	4.10	1.68	3.39	3.9	2.53	1.75
Range of x- coordination	36	30	28	50	46	39	18
Range of y- coordination	27	35	11	24	22	14	11
Range of Actual HR (bpm)	7.85	6.32	8.34	10.37	13.08	3.62	9.24
SD of Actual HR (bpm)	2.13	1.55	2.02	2.63	3.74	1.00	2.68
Results Using ICA							
Selected Component	1st	1st	1st	3rd	2nd	1st	1st
Mean bias (bpm)	2.82	-4.07	11.95	-1.59	2.55	-10.30	10.42
SD of bias (bpm)	4.47	1.94	6.07	20.27	5.70	1.50	9.04
Max bias (bpm)	14.12	0.12	27.22	25.40	14.17	-7.93	31.06
Min bias (bpm)	-12.95	-8.89	-0.38	-49.13	-15.47	-18.43	-2.08
RMSE	5.30	4.52	13.44	20.34	6.25	10.47	13.84

However, exploring the whole experiment as one dataset showed that the 3rd component produced the RMSE of 35.23 which was the best value among the components and the MPA method. On the other hand the kNN technique dramatically increased the performance of the estimation. Figure 4-4 illustrates an example of the improvement of ML techniques over the ICA in measuring HR.

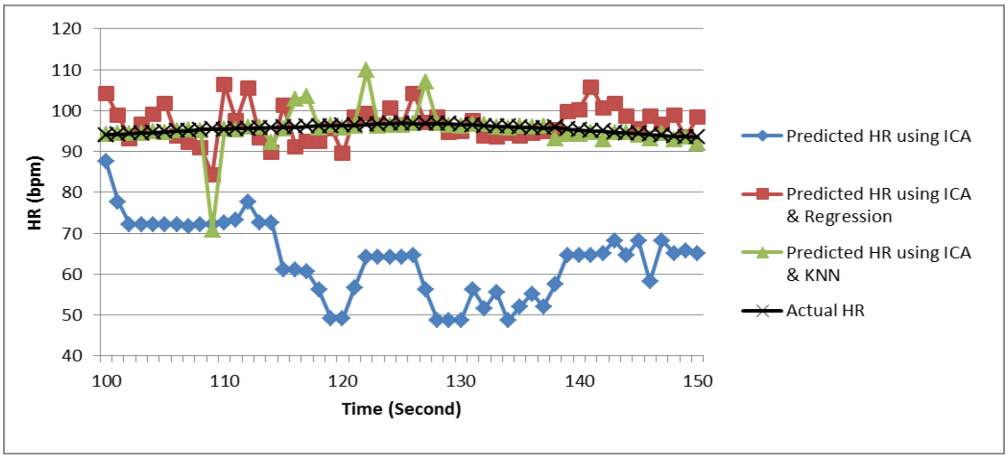


Figure 4-4: Actual HR changes during the 3rd experiment (Indoor exercising) extracted from ECG.

The Bland-Altman analysis showed that the best agreement between actual HR measured by ECG and the proposed methods was achieved by the ICA and kNN method. This method reduced the mean bias from -24.37 bpm to 0.05 bpm with 95% LoA -8.20 to 8.76. Table 4-8 summarizes the results of HR detection in the third study. These results represented the robustness of the proposed method in a new condition which includes lots of user motions and a large range of HR changes (from 69.5 bpm to 130.79 bpm).

Table 4.8: The descriptive statistics of HR estimation using ICA method and ML techniques during indoor exercising (third experiment)

	Third Experiment		
	ICA	**ICA + kNN**	**ICA + Regression**
Selected component	3rd	-	-
Mean bias (bpm)	-24.37	-0.28	0.05
SD of bias (bpm)	25.54	4.33	13.70
RMSE	35.31	4.33	13.69
Corr. Coefficient	0.53^{***}	0.97^{***}	0.58^{***}
Mean absolute error	28.53	1.41	11.24
Relative absolute error	2.09	0.10	0.82
Root relative squared error	4.37	0.25	0.81

***: (p<0.001);

4.4 Conclusion

We have evaluated a method for remote HR measuring in three applications: a controlled laboratory task, a naturalistic HCI and an indoor cycling exercise. This study evaluated Poh et al.'s method and showed the feasibility of their methodology to measure HR at rest. Their seminal work is one of the successful attempts at remote physiological sensing. However their proposed method did not show positive results in naturalistic HCI and indoor exercise situations. The study analyses the problems caused by unwanted movements and the wide dynamic range of HR which are common during real world measurement. We have addressed these issues by building and training specific models for each participant using ML techniques. The results suggest that the kNN based technique outperforms other approaches (manual or computational) that try to select the best independent component for HR estimation. On average the mean of absolute errors in HR estimation is reduced to 0.68 bpm by applying kNN technique to the ICA outputs among 10 participants in the HCI scenario. The kNN technique also improved the accuracy of the HR estimation over the former method in indoor exercising conditions.

Although the accuracy of estimation using the proposed approach is increased, in some applications (e.g. emergency severity index (ESI) triage) the achieved accuracy is not acceptable and more improvements are needed to implement a more accurate estimation. Another limitation might be the objection to drawing major generalizations from the small number of subjects studied in the third study. It should be noted that the purpose of the third study was to measure the impact of a larger dynamic range on the measurements. The lighting conditions and other environmental variables were kept the same. Recruiting more participants for generalizing our findings from the third study is considered for possible future work. However building a user-independent model to yield reliable results when presented with new users without the need for retraining is one of our concerns for future work.

Chapter 5. Study 2: Detecting non-basic emotion in HCI (IAPS)

Summary

In this chapter, the second part of the experiment introduced in Chapters 3 is analysed. In this experiment, 60 images were used as emotional stimulus. For each participant, 60 video segments were recorded and analysed here. The results of two types of concurrent self-report are reported first. Then the result of applying feature extraction methods are presented. The classification results are also analysed at the end of this section. User-dependent, gender-specific, and user independent analysis are evaluated (Monkaresi, Calvo, & Hussain, 2014).

5.1 Introduction

Several methods, techniques, and devices have been proposed in the past for affect detection. Some of them relied on single modalities like facial expression, voice, and physiological signals that were successful for detecting even complex affective states (Calvo & D'Mello, 2010). However, multimodal affect detection techniques are becoming increasingly popular due to their reliability and performance in detecting complex affective states (D'Mello & Graesser, 2010; Hussain & Calvo, 2011; Pantic & Rothkrantz, 2003; Soleymani, Pantic, et al., 2012). Naturally, humans use several modalities when they are interacting with each other. Each modality (face, voice, gesture, physiology, etc.) has a unique characteristic of an affective state and considering more modalities can increase reliability and accuracy of affect interpretation.

Physiology is a prominent modality that has been used for affect detection because it is suitable for reflecting inner feelings, is robust against social marking, and offers good time resolution. It has also been used in multimodal affect detection approaches (Hussain et al., 2011; Soleymani, Pantic, et al., 2012).

Normally, physiological sensors need to be attached to the human body, which might be intrusive and make the application hard to adopt. Wearable sensors and devices were proposed to reduce the hardships of setting up the traditional sensors. Among the current methods of measuring physiological signals, contact-less and remote methods are more desirable. These methods are easy to adopt and much cheaper than traditional devices (Poh et al., 2010). A remote, contactless sensor can monitor several subjects at the same time.

In this chapter, we utilize the proposed method in Section 3-3-3 to measure HR remotely and use it for affect detection in combination with facial expression features. A dynamic approach has been used to extract facial expression features based on local binary patterns in three orthogonal planes (LBPTOP) (See Section 3-3-1). As for the third channel, the Kinect face tracker engine (See Section 3-3-2) was used to extract facial geometric-based features. Then a fusion model was utilized to classify affective states using these three channels. In the following section, we briefly explain two preliminary experiments for evaluating the possibility of using geometric-based in combination with physiological channels for affect detection.

5.2 Preliminary studies

Before this study, two preliminary studies were undertaken to evaluate the possibility of detecting non-basic affective states using a combination of geometric-based features from the face and physiological signals. We used the dataset recorded by two colleagues from our group (Learning & Affect Technologies Engineering, LATTE). The dataset included the data collected from 20 participants. Ninety images from the IAPS dataset were selected and used for triggering participants' affective states. The annotations were recorded in 2D space that contained three levels (low, medium, and high) for valence and three levels for arousal. A video of frontal face and four physiological signals were recorded while the participants viewed the images. The video was recorded using a webcam. Four physiological signals, including electrocardiogram (ECG), electromyogram (EMG), respiration, and galvanic skin response (GSR) were recorded using a BIOPAC MP150 system with AcqKnowledge software at 1000 samples per second for all channels. The achieved Kappa measures for each method are shown in Table 5-1.

Table 5.1: Average kappa measures for detecting three levels of normative Valence (V) and Arousal (A) in preliminary studies

Preliminary studies		Kappa measures			
		User-Dependent Model	General model	Gender Specific	
				Male	Female
AVI (Monkaresi, Calvo, et al., 2012)	kNN Classifier (k=1), one modality (head movements)	V:0.37 A:0.30	V:0.14 A:0.16	–	–
SMC (Monkaresi, Hussain, et al., 2012b)	Vote classifier, Fusion model (Head movements and physiological signals)	V:0.47 A:0.38	V:0.22 A:0.19	V: 0.32 A: 0.16	V: 0.24 A: 0.22

In the first study (Monkaresi, Calvo, et al., 2012) geometric-based features that indicated the position of the head were extracted and used for non-basic affect detection. Three classifiers (kNN, linear SVM, and Bayesian Network) were used for affect classification, and the kNN classifier showed the better performance compared with the others. The results showed that the user-independent model could not be crated based on extracted head movement features for valence and arousal detection. Other features from the face needed to be added to the model to increase the accuracy of affect detection. However the user-dependent models yielded moderate accuracy for discriminating between the three levels of valence and arousal.

In the second study (Monkaresi, Hussain, et al., 2012b), a fusion model was proposed to combine the physiological data head movement features for affect prediction. A combination of classifiers (Vote classifier with the average probability rule) was used for classification. On average, the vote classifier showed the better performance compared with the three classifiers used in first preliminary study. The results also showed that the fusion model outperformed the physiological channels for detecting valence and arousal using user-dependent and gender-specific models. According to the Kappa scores, the fusion model obtained better accuracies for valence detection compared with the video channel. However, the fusion model did not show any improvement over the video channel for detecting arousal levels.

5.3 Participants

For this study, 23 undergraduate/postgraduate engineering students from the University of Sydney were recruited for the experiments. The participants' ages ranged from 20 to 60 years (M = 34 years, SD = 11) and there were 14 males and 9 females. There were 5 Asians and 17 Whites, and 6 participants wore eyeglasses. We advertised our experiment by circulating the flyer through the University of Sydney's student newsletter and the Centre of the Research and Innovation website[4]. An example of the flyer is presented in Appendix B. The University of Sydney's Human Ethics Research Committee approved the study prior to data collection. The participants signed an informed consent prior to the study. The approved consent form is available in Appendix C.

All participants fulfilled both parts of the experiment (writing session and IAPS session). There was a synchronization problem between video segments and self-reports for one participant and the participant was ignored for feature extraction and affect classification. Therefore, the classification results are only reported for the recorded data of 22 participants.

5.3.1 Sensors and experiment setup

The experiment was conducted indoors with a varying amount of ambient sunlight entering through windows in combination with normal artificial fluorescent light. Participants were asked to sit in front of a computer and interact normally while a Microsoft Kinect sensor (PC version) recorded their video. All videos were recorded in colour (24-bit RGB with 3

[4] http://sydney.edu.au/research/involved/volunteer.shtml

channels, 8 bits/channel) at 30 frames per seconds (fps) with pixel resolution of 640 × 480 pixels and saved in AVI format. Two physiological signals, ECG and respiration, were also recorded using a BIOPAC MP150 system with AcqKnowledge (v. 3.8.2) software. The acquisition sampling rate was 250 Hz for both channels. Three electrodes were placed on the participant's body to record the ECG signals: two electrodes were placed on their arms and the ground electrode was placed on their ankle. Instead of the wrists, the electrodes were placed on the arms to reduce the noises that might be introduced by hand movement during typing. Figure 5-1 shows the experimental setup was used in the studies reported in this chapter and Chapter 6.

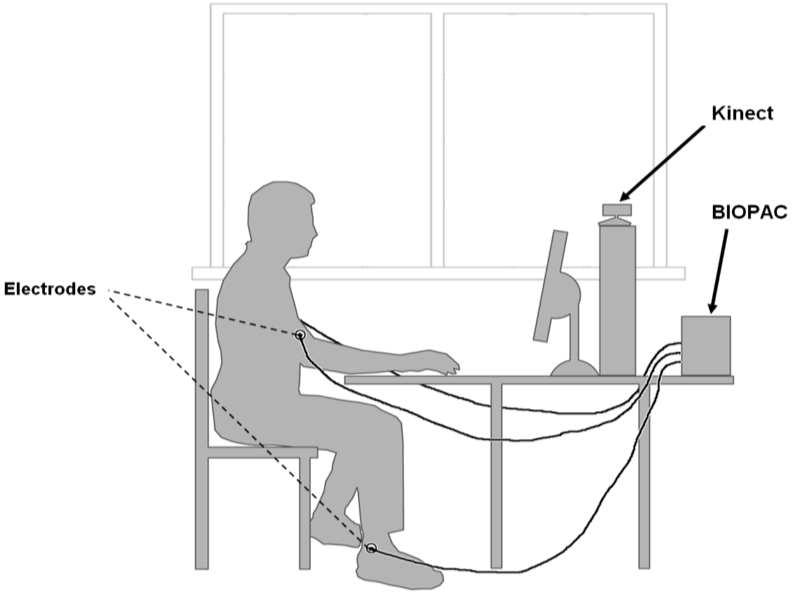


Figure 5-1: Experiment setup for Studies 2 (IAPS images) and 3 (Writing)

5.4 Analysis of self-reports

Two types of self-reports were recorded for this experiment: dimensional (valence/arousal) and categorical self-reports. These annotations were extracted concurrently during the experiment. All 23 participants completed the experiment and finally, 1380 instances were produced (23 participants × 60 video segments). The statistics of each reported score and emotional category were reported.

5.4.1 Valence/arousal

Valence and arousal levels were recorded using a standard 1–9 rating scale. Since we tried to select the images from different rating scales, we expected that there was an adequate number of labels for each rating scale. Figure 5-2 shows the distribution of reported scores in this experiment. According to Figure 5-2, a normal distribution of reported scores can be observed.

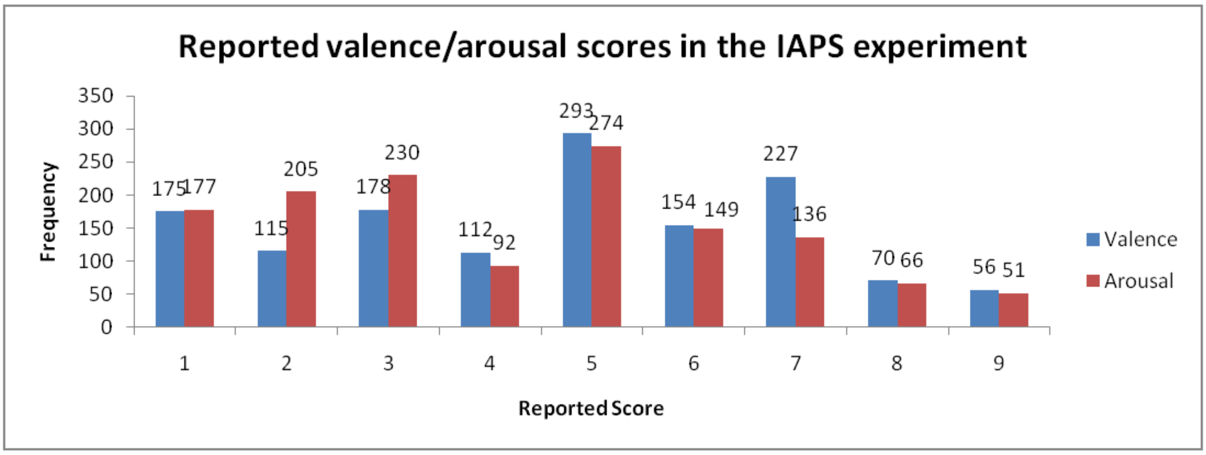


Figure 5-2: Distribution of reported scores in concurrent self-reports (IAPS experiment)

To address the individual variations in ratings, the reported scores were standardized (converted to z-scores) for each participant. After standardization, the mean of the participant ratings was equal to zero. The z-score was negative when the raw score as below the mean, and it was positive when above. Then, the positive z-scores were considered as “High Arousal” (or “Positive Valence”), and the negative z-scores were considered as “Low Arousal” (or “Negative Valence”). Accordingly, the classification task discriminates between two levels in valence and arousal dimension. Figure 5-3 shows the percentage of reported values in each class after this conversion. Figure 5-3 shows an almost balanced class distribution in valence and arousal dimensions, which is desirable for training a classifier.

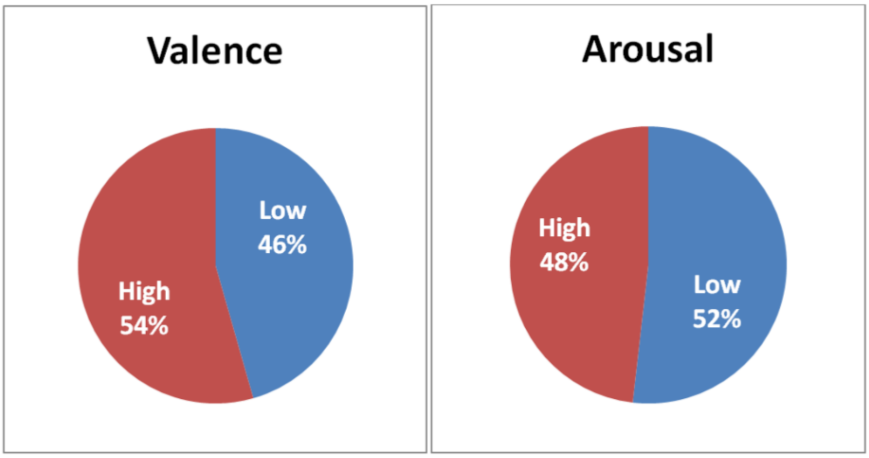


Figure 5-3: The percentage of reported valence and arousal after standardization and grouping into two levels (IAPS experiment)

5.4.2 Categories

One thousand, three-hundred and eighty emotional categories were also concurrently reported by all 23 participants. Figure 5-4 presents the percentage of each reported category using a pie chart (Calm: 164, Relaxed: 260, Bored: 111, Annoyed: 106, Glad: 102, Content: 37, Delighted: 63, Excited: 99, Depressed: 97, Gloomy: 61, Afraid: 90, Angry: 60, Others: 130). *Relaxed* and *Calm* were reported more frequently in the IAPS experiment. *Content*, *Gloomy*, and *Angry* were reported less than other categories.

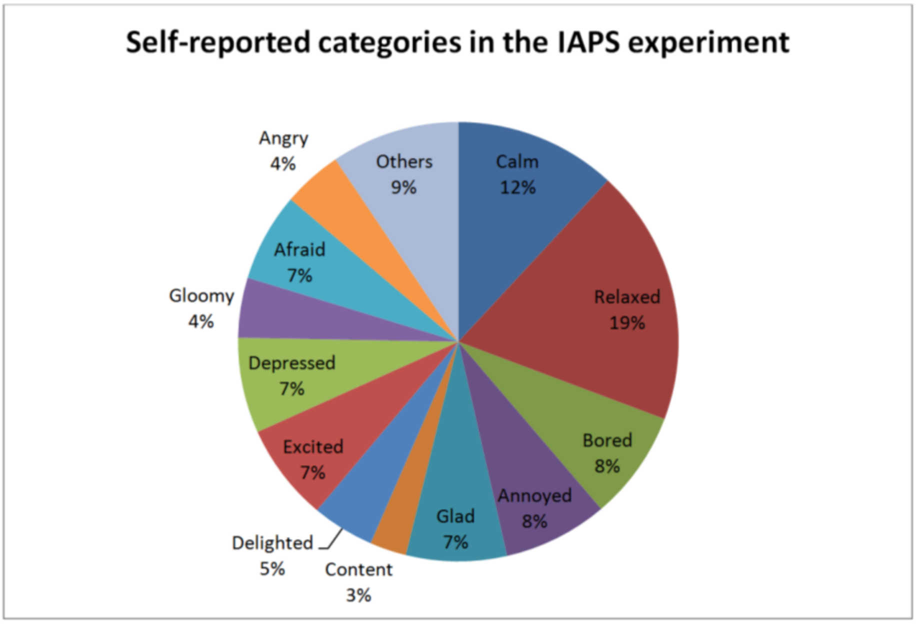


Figure 5-4: The percentage of each reported emotional category in the concurrent self-report (IAPS experiment)

5.4.3 Categorical vs. dimensional affect

The relations between dimensional scores and emotional categories were explored. Standardized scores were used for this reason. The average valence and arousal z-scores were calculated for each category across all labels. The position of each category in 2D space (valence vs. arousal) is plotted in Figure 5-5. The result almost supports the well-known circumplex of affect reported by Russell in 1980. *Excited* had the highest level of arousal and a high level of valence, whilst *Angry* had a high arousal with the most negative value of valence.

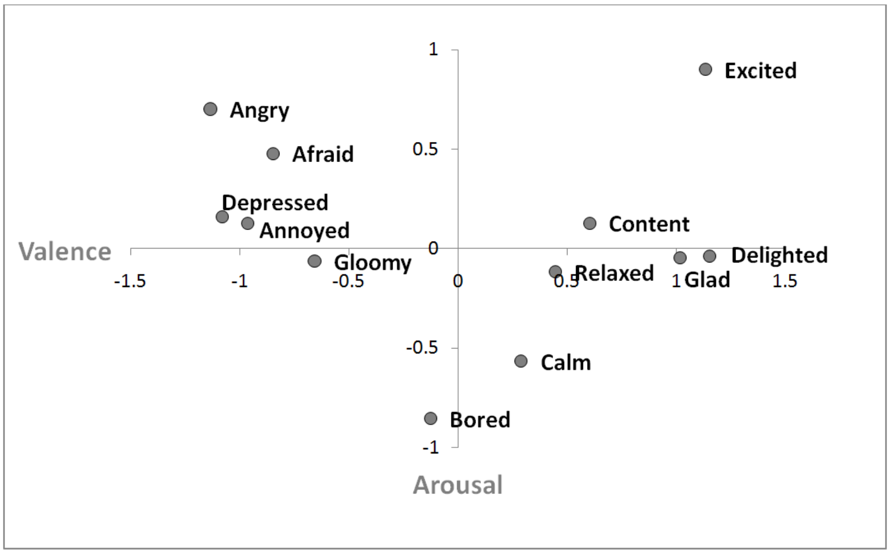


Figure 5-5: Mapping the reported categories on the 2D space of affect (valence/arousal)

5.5 Feature extraction

Three types of video-based features were extracted and analysed in this study. For each video segment, 84 features were extracted by the Kinect Face Tracker method. Two thousand, three hundred, and four LBPTOP features were extracted for each video segment as described in Chapter 3. The HR signal was also extracted using the method proposed in Chapter 4. Each video segment last 10 seconds and for each second, there is an estimate for HR. Seven statistical features were extracted from the HR estimations for each video segment. Altogether, 2395 features were extracted and then synchronized with corresponding annotations (normative ratings and concurrent self-reports). The following section reports the classification accuracies for detecting valence and arousal. The performance of each modality in discriminating between two levels of valence and arousal was also explored and reported.

Two participants were too close to the screen while viewing the stimuli and facial features could not be extracted. The eye-related features could not be extracted for three participants due to occlusion caused by eyeglasses. Therefore, the classification results are reported based on recorded data from 17 participants.

5.6 Classification results

This section reports the classification results for detecting degrees of valence and arousal using three set of features (channels) and a feature level fusion of these three channels. The results are reported in three sub-sections: user-dependent, gender-specific, and general (user-independent) models. For each model, six classification tasks were performed for detecting six types of affect representations. Valence and arousal was considered independently in four representations and in two representations, a combination of valence and arousal was considered. Table 5-2 describes these six types of affect representation, which have been used in the following classification.

Table 5.2: List of six classification tasks used for each model

Abbreviation	Source	Affect	# of classes	Name of classes
selfVal	Concurrent Self-report, Dimensional	Valence	2	Low-valence, High-valence
selfAro	Concurrent Self-report, Dimensional	Arousal	2	Low-arousal, High-arousal
selfDim	Concurrent Self-report, Categorical	Valence × Arousal	5	Low-val_low-aro, Low-val_high-aro, High-val_low-aro, High-val_high-aro, Other
normVal	Normative rating	Valence	2	Low-valence, High-valence
normAro	Normative rating	Arousal	2	Low-arousal, High-arousal
normDim	Normative rating	Valence × Arousal	4	Low-val_low-aro, Low-val_high-aro, High-val_low-aro, High-val_high-aro,

5.6.1 User-dependent models

In user-dependent analysis, 17 specific models were trained and tested for each participant. Figure 5-6 presents the average Kappa scores for classifying affective states using three separate channels (HR, FT, LBP) along with the fusion model. The performance of each channel in detecting the six types of affective states using user-dependent models is discussed here.

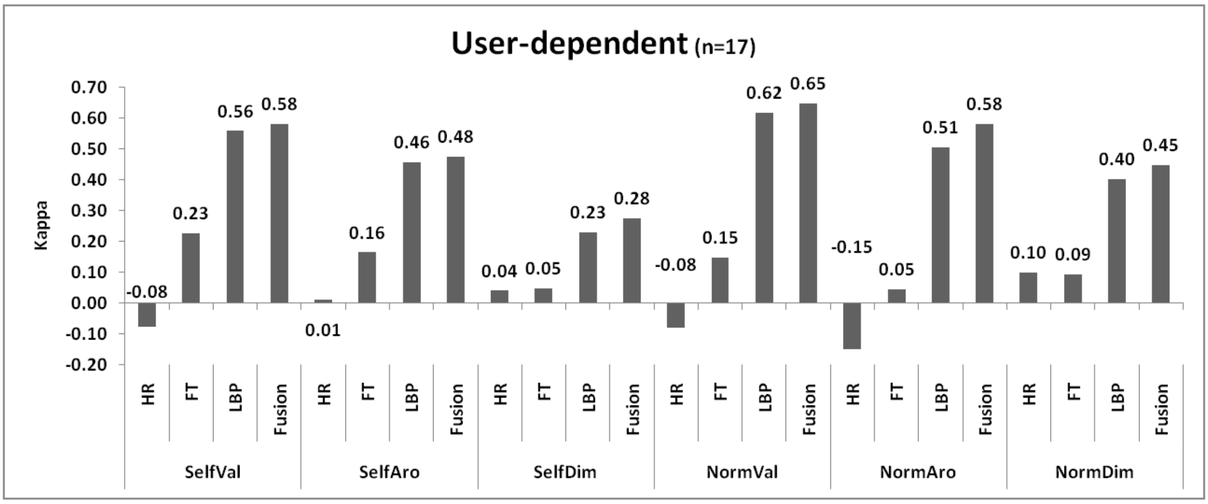


Figure 5-6: The average Kappa scores for classifying affective states using 10-fold cross validation approach (user-dependent models)

__Fusion model:__ For all classes, on average fusion models achieved the best results with a reasonable accuracy compared with each individual channel. The best result (Kappa = 0.65) was achieved by the fusion model for classifying two levels of normative valence. The improvements of fusion model over the HR and FT channels were statistically significant in all six cases as specified by paired samples *t* tests ($p < 0.05$). The paired samples *t* tests also indicated that the fusion model significantly improved the Kappa scores over the LBP channel for detecting *selfDim* (t(16) = –3.00, $p < 0.05$) and *normAro* (t(16) = –2.59, $p < 0.05$). However, the improvements of the fusion model over the LBP channel were not significant for detecting *selfVal* (t(16) = –1.00, *p* = 0.33), *selfAro* (t(16) = –0.61, *p* = 0.55), *normVal* (t(16) = –0.70, *p* = 0.49) and *normDim* (t(16) = 1.71.00, *p* = 0.11). In general, these results suggest that adding the FT and HR to the LBP channel could improve the accuracy of affect detection.

<u>HR:</u> The HR channel showed a weak performance for detecting *normDim* average Kappa measure of 0.10. The HR modality obtained an average Kappa measure of 0.04 for detecting self-reported combination of valence and arousal (*selfDim*). This channel did not show any improvement over the chance estimations as indicated by negative values of Kappa scores. HR obtained the maximum Kappa measure of 0.47 for detecting *selfAro* using HR channel.

<u>FT:</u> The Kinect face tracker features set (FT) was successful in discriminating between negative and positive self-reported valences, *selfVal* (average Kappa = 0.29). Even though it achieved positive Kappa scores for classifying other types of affective state, the average performance was not accurate enough. On the other, the FT channel showed an excellent performance for some participants. For example, this channel achieved the Kappa measure of 0.93 for detecting normative valence (*normVal*). Compared with other channels, the FT channel represented better performance for classifying self-reported labels than normative ratings.

<u>LBPTOP</u>: The LBP channel obtained the best Kappa scores among the three channels for detecting different types of affect representations. Adding HR and FT features did not provide a big improvement over the accuracies achieved by the LBP channel. Except for the *selfDim*, the LBP channel showed a good performance for affect detection, with average Kappa measures ranging from 0.40 to 0.68. An excellent agreement between classified and actual labels was achieved for normative valence and arousal, with average Kappa measures of 0.62 and 0.51, respectively. The results for detecting self-reported valence (average Kappa = 0.56) and arousal (average Kappa = 0.46) were also promising using the LBP channel. The results also demonstrated that the LBP channel performs significantly better than other individual channels.

Compare Affective states: According to Figure 5-6, the performance of the system for detecting valence was better than arousal and the combination of valence and arousal (*selfDim* and *normDim*). From the figures, we discovered that considering valence and arousal separately could produce a more accurate prediction. For example, the fusion model predicted normative valence and arousal with Kappa measures of 0.65 and 0.58, respectively, whilst the Kappa measure for detecting *normDim* was 0.45.

On the other hand, the results show that compared with self-reported labels, normative ratings were classified more accurately. This can demonstrate the reliability of the IAPS labels, which is expected because an adequate number of participants (around 100) were used to produce those labels. However, in our study, each subject might have had his/her own understanding about the rating scales and subjective variations might be one reason for the reduction in affect detection accuracies.

5.6.2 Gender-specific models

For this analysis, we separated our dataset into two parts, with one part containing only male participants' data (*n* = 10) and the other part only female participants' data (*n* = 7). Then, data from individual participants were standardized (converted to z-scores) to address individual variations of head behaviour and physiological differences. We built and trained specific models for each of the datasets to compare the performance of gender-specific models. A leave-one-out cross-validation approach was used to evaluate these models, which was described in Section 3-4. The Kappa measure for detecting affective states in the males and females models are shown in Figures 5-7 and 5-8 respectively.

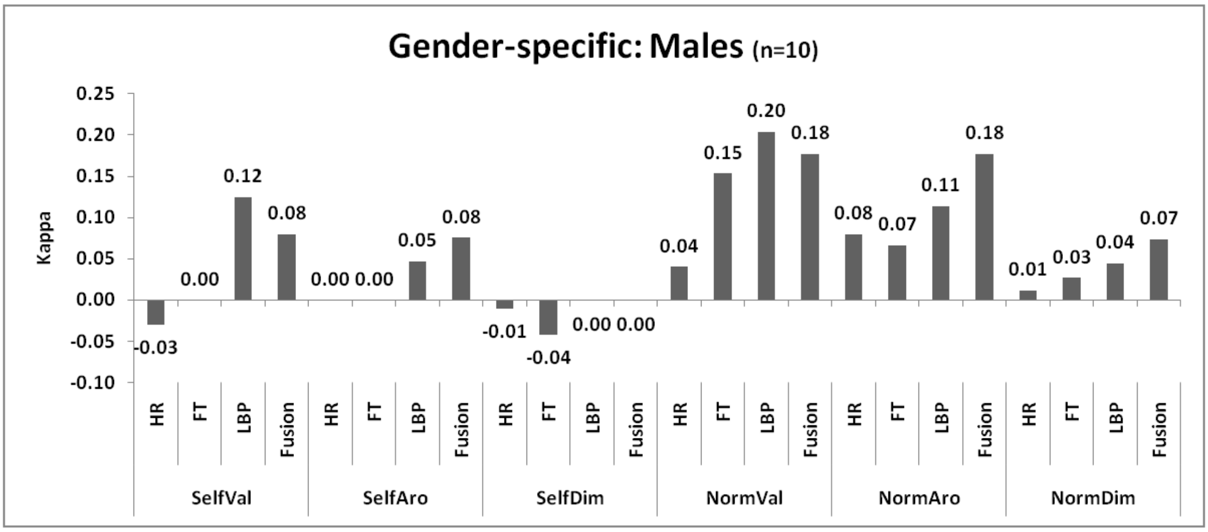


Figure 5-7: The Kappa measures for classifying affective states using leave-one-out cross validation approach (Gender-specific: males)

Male model: As Figure 5-7 shows, no reasonable accuracy was achieved for classifying self-reported affective states. The best result among self-reports was obtained by the LBP channel, with a Kappa measure of 0.12 for detecting valence. Subjective variations across the male participants' ratings might be the main reason for the weak performance of the system for detecting self-reported affect. However, the achieved results for detecting normative affective states were more promising. All channels achieved positive Kappa measures. Again, the LBP channel obtained the best result for classifying normative valence for male participants. The fusion of the FT, HR, and LBP channels improved the Kappa measures for detecting normative arousal and normative combination (of valence and arousal) by 0.07 and 0.03, respectively, whereas the fusion of these channels failed to improve the accuracy of normative valence.

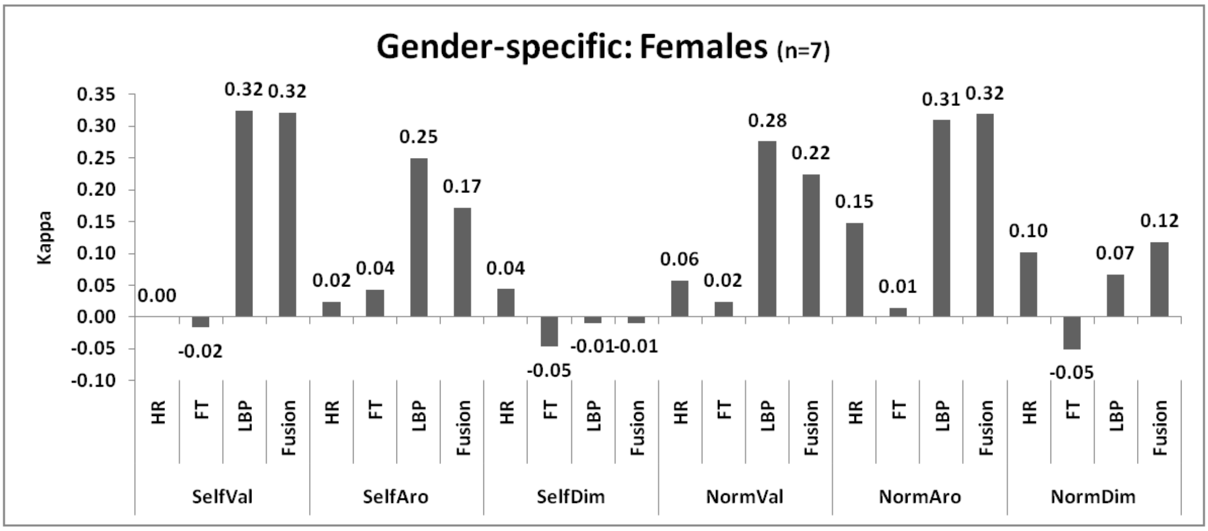


Figure 5-8: The Kappa measures for classifying affective states using leave-one-out cross validation approach (Gender-specific: females)

<u>Female model</u>: In the females gender-specific model (Figure 5-8), the best performance was achieved by the LBP channel for detecting self-reported valence (Kappa measure = 0.32). The fusion model was also successful in discriminating between different degrees of self-reported valence (Kappa measure = 0.32) and normative arousal (Kappa measure = 0.032). The best Kappa score for the HR channel was obtained for detecting normative arousal (Kappa measure = 0.15). The HR channel also achieved the Kappa measure of 0.10, which represents a weak performance for detecting four classes of affective states in valence × arousal space (*NormDim*).

Except for the *SelfVal*, *NormAro*, and *NormDim*, the fusion model failed to improve the accuracy of affect detection. For the rest of the affect types, the fusion model improved the accuracy of affect detection slightly.

Overall, the female gender-specific model showed the higher Kappa measures compared with the male gender-specific model. This finding suggests that females express their emotions through facial expressions more than males. The higher Kappa values in self-reported affective states could reflect the quality of the self-reports, which might be because of more effort from the females in filling out the questionnaires.

5.6.3 General model

To build a user-independent model, data from individual participants were first standardized and then combined to yield one large data set with 1,020 instances. The Kappa measures were calculated based on leave-one-participant-out cross validation approach. Figure 5-9 shows the Kappa measures for detecting the affective states using the general model.

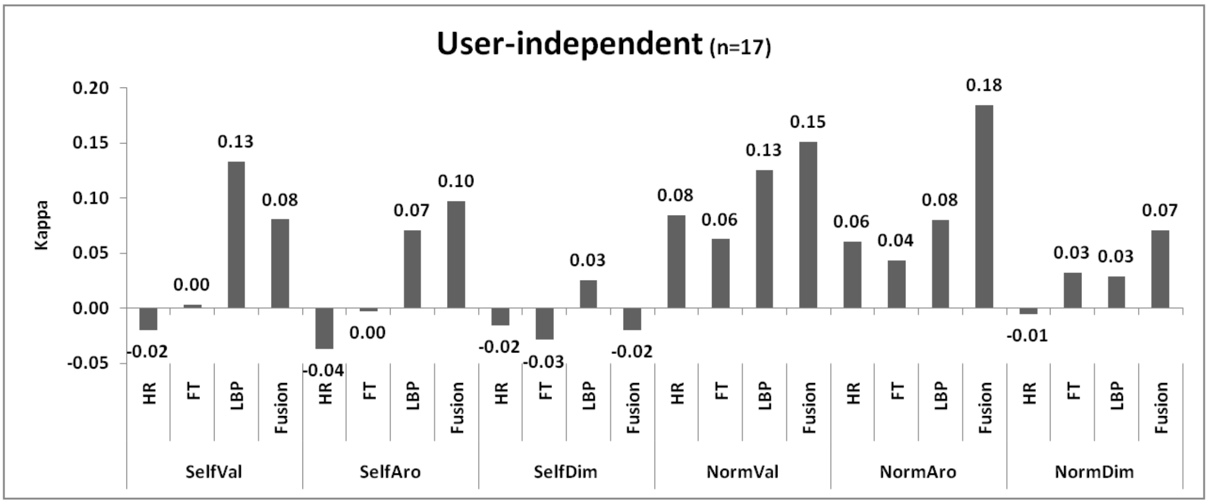


Figure 5-9: The kappa measures for classifying affective states using leave-one-out cross validation approach (User-independent model)

As expected, the results for the self-reported affective states were not promising enough. The best result for detecting self-reported valence was achieved by the LBP channel with a Kappa measure of 0.13. The fusion model improved the accuracy of self-reported arousal detection by 0.10.

However, in the normative ratings, the fusion model improved the results achieved by all other channels successfully. The Kappa measures were increased from 0.13 and 0.08 to 0.15 and 0.18 for detecting normative valence and arousal, respectively. A supper additive improvement was also achieved by the fusion model for discriminating between four classes in valence and arousal space (*NormDim*). These improvements showed the success of the fusing the HR, FT, and LBP channels in affect detection.

The HR channel also produced positive Kappa measures for classifying normative valence and arousal.

5.7 Feature selection

To have a better understanding about the contribution of each channel in the affect classification task, we report the selected features here. In this section, we only discuss the selected features for the user-independent model as shown in Table 5-3. In this table, the name of LBPTOP features starts with *P* character followed by the pattern number. The name of HR features start with *HR*, and the rest of the features belong to the FT features.

According to Table 5-3, the LBPTOP features (P) contributed in classification of all types of affective states. The FT features also had a good contribution for affect detection. The interesting results were about the HR features. The results indicated that the HR related features contributed in detecting four out of six types of affect. Among seven HR-related features, *HR-median* was the most common feature contributing to affect detection.

Table 5.3: List of selected features from the fusion model (user-independent analysis)

Selected features from Fusion model					
SelfVal	**SelfAro**	**SelfDim**	**NormVal**	**NormAro**	**NormDim**
ANU2-std	ANU4-max	Ty-mean	ANU5-max	Tz-std	Tz-mean
ANU5-min	ANU4-range	P803	Rx-mean	**HR-mean**	**HR-median**
Tz-mean	**HR-mean**	P814	Ty-min	**HR-median**	**HR-max**
Tz-max	**HR-median**	P835	Tz-mean	**HR-min**	**HR-range**
P107	P70	P853	Tz-median	**HR-max**	P17
P114	P427	P886	**HR-median**	**HR-range**	P348
P216	P1877	P983	**HR-max**	P17	
P230	P1955	P1386	P62	P48	
P773	P1968	P1558	P72	P201	
P795	P1973	P1620	P165	P348	
P820		P1707	P216	P352	
P931			P764	P463	
P942			P894	P563	
P984			P1316	P974	
P1337			P1617	P1001	
P1341			P1650	P1373	
P1684			P1658	P2022	
			P1684	P2038	
			P1739	P2091	
			P1773		
			P1777		
			P2221		

The HR features had their maximum contribution in detecting normative arousal where the fusion model achieved the best results (Kappa measure = 0.18). Five out of seven HR features were selected for the normative arousal detection. In addition, 50% of selected features for detecting *NormDim* were also from the HR features. Overall, these results showed that the HR changes play an important role for affect detection in HCI applications.

ANU5 and Tz-related features appeared among the selected features for detecting *SelfVal* and *NormVal*. This suggested that these features could be the good indicators for valence detection. ANU5 indicates the level of "Outer Brow Raiser" and Tz indicates the motion of the head in the Z axis. Tz also contributed in detecting other types of normative affective states. This finding suggested that moving the head in the Z axis can be a good indicator for affect detection. When you have a positive feeling about an image you are watching, you probably move toward the screen to watch it precisely. Within the face tracker (FT) features, ANU4, which specifies the "Lip corner depressor", was the best indicator for detecting *SelfAro*.

Figures 5-10 and 5-11 show the contributions of different LBPTOP features more precisely. These figures divide the selected features into six categories from two perspectives. In the first perspective (Figure 5-10), the selected features were divided based on two regions: eyes and mouth. In the second perspective (Figure 5-11), the selected features were divided according to the planes set that they belong to. The XY-related features were considered as appearance-based features, and features extracted from XT and YT planes were considered as the motion-based features.

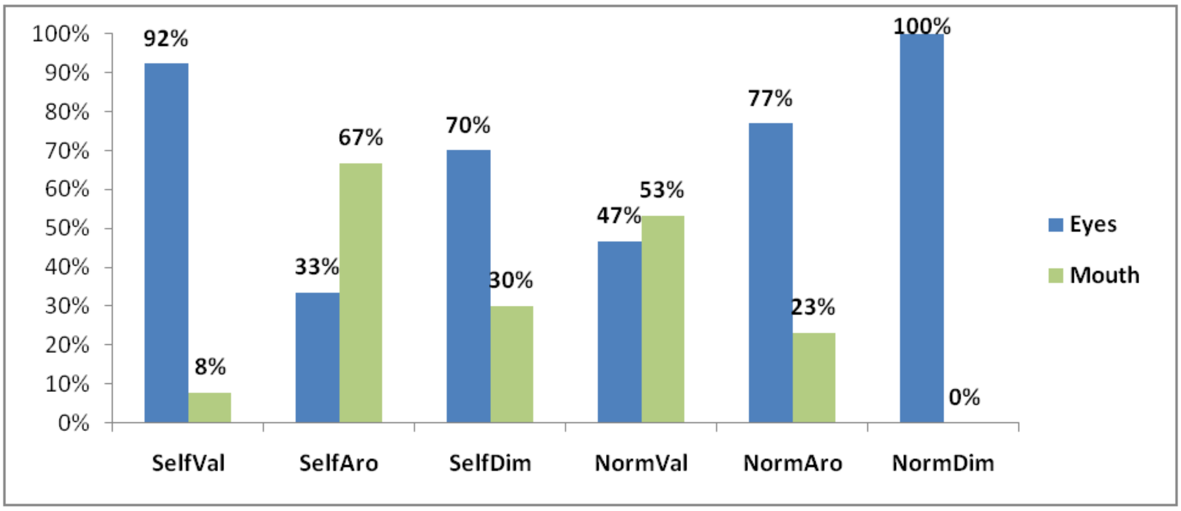


Figure 5-10: The percentage of contributions of the facial objects (eyes and mouth) in selected LBPTOP features for affect detection using the fusion model (user-independent analysis)

In general, among the LBPTOP features, eyes-related features contributed more in the affect classification. Seventy-five percent of selected LBPTOP features for detecting normative affective states were selected from the eyes-related features. However, the distribution of mouth-related features among the selected LBPTOP features for classifying *SelfAro* and

NormVal was more than the eyes-related features. On the other hand, the mouth-related features did not contribute in detecting *normDim*.

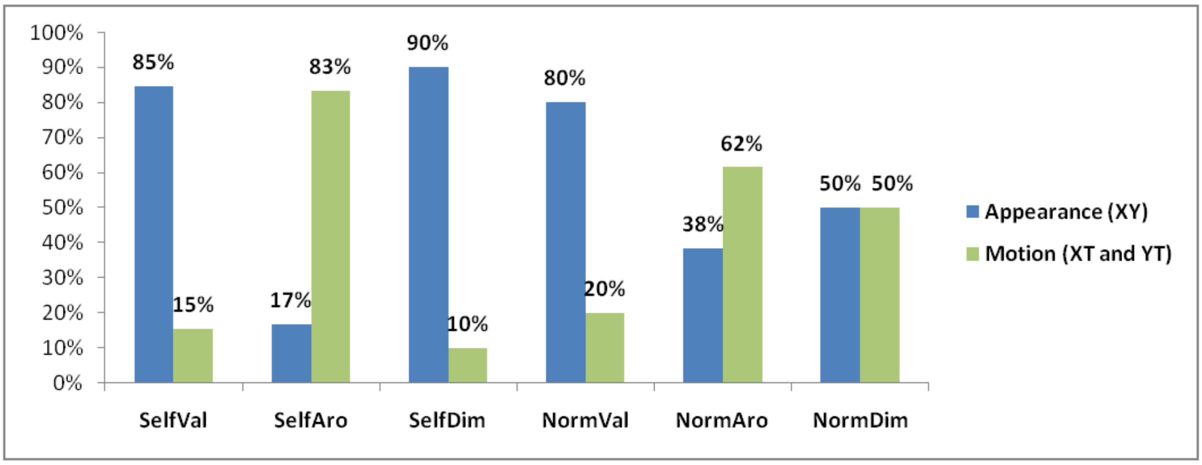


Figure 5-11: The percentage of contributions of appearance-based and motion-based LBPTOP features for affect detection using the fusion model (user-independent analysis)

According to Figure 5-11, the appearance-based features were more involved for detecting self-reported and normative valence compared with the motion-based features. On the other hand, the motion-based features contributed more in detecting self-reported and normative arousal than appearance-based features. There was the same contribution of these two types of LBPTOP features for detecting *NormDim* whilst the appearance-based features contributed more than motion-based features for detecting *Self-Dim*. Overall, the results suggested that the appearance-based features were more useful for detecting valence and the motion-based features were useful for detecting arousal. It should be mentioned that both types of features are essential for affect detection.

5.8 Conclusion

This chapter introduced a new fusion model for affect detection. In this model HR-related features extracted from facial videos were combined with geometric-based and appearance-based facial features. The results showed that combining these HR features with other facial expression features (e.g., LBPTOP features) could improve the accuracy of affect detection using the normative rating. The HR channel was more successful in detecting normative arousal for female participants (Kappa measure = 0.15). The fusion model also showed reasonable accuracy for detecting affect (normative rating) in user-independent analysis. However, the dynamic features (LBPTOP) achieved the best performance in discriminating between different levels of each type of affective states.

This study demonstrates the feasibility of using contact-less physiological signal measurements for affect detection even though the improvement was slight. Replacement of traditional physiological sensors with a camera could significantly increase the usability of an affect detection system. The approach in this chapter is the first attempt in using remote physiological measurement with other channels for affect detection and improvements needs to be done. Extracting more physiological signals, such as inter-beats intervals and respiration rates using the video-based method and adding them to the fusion model could be considered as future works.

Three types of analysis have been evaluated in this chapter: user-dependent, gender-specific, and a general model. In most of the analysis, fusion models achieved the best results. As expected, the user-dependent models obtained the best results compared to other two analyses. Evaluating the gender-specific models also suggested that developing a separate model for each gender could increase the accuracy of affect detection.

At the end of this chapter, a comparison between selected features has been provided, which gives a better understanding about the relation between each affective state and each set of extracted features.

Chapter 6. Study 3: Detecting engagement and affect during writing

Summary

In this chapter, we report the results of a study into detecting user's engagement level and his/her affective states during computer-based writing session using the methods introduced in Chapter 3. This chapter starts with providing the information about the participants and materials followed by the self-reports analysis. The extracted features using the presented methods and their ability in detecting engagement and affect are evaluated through analysis of feature selection results. Then, the classification accuracies are calculated and validated with 10-fold segment-level cross validation for user-dependent models and leave-one-subject-out validation for user-independent models.

6.1 Introduction

As mentioned in the literature review, most previous research on affect detection focused on detecting basic emotions (D'Mello & Calvo, 2013; Zhihong Zeng et al., 2009). More recently, researchers in the this field have shifted towards recognizing complex mental states and particularly, attention and engagement (Bohus & Horvitz, 2009; Grafsgaard et al., 2013; McDaniel et al., 2007; Nakano & Ishii, 2010). Engagement is an important mental state related to productivity and learning (Christenson, Reschly, & Wylie, 2012b; Kahn, 1990b). For example by measuring the level of student's engagement in a classroom, the teacher can change the teaching method to promote or hinder engagement. Nowadays, computer-based activities are more popular in workplaces and educational environments. Measuring task engagement during human computer interaction needs more specific considerations. Peters et al. (2009) discussed different aspects of engagement that need to be considered in the HCI applications. Several researchers have defined three types of engagement (Fredricks, Blumenfeld, & Paris, 2004; Peters et al., 2009): behavioural engagement, which can be observed by someone; emotional engagement, which can be assessed by measuring emotional reactions to a task; and cognitive engagement, which is an internal mental activity and hard to measure.

In this chapter, we use our proposed methodology to detect engagement and affective states during the writing activity. Writing is one of the most common activities in the workplace and in educational environments. Enabling computer-based writing tools to recognize a user's engagement and affective state can help users enjoy their writing activities. The possible relation between engagement and affective states (valence and arousal) has been explored. In addition, the impact of feedbacks and casual interventions on participants' affective states and engagement has been analysed.

6.2 Participants

Twenty-three undergraduate/postgraduate engineering students from the University of Sydney were recruited for the experiments. The participants' ages ranged from 20 to 60 years (M = 34 years, SD = 11), and there were 14 males and 9 females. We advertised our experiment by circulating the flyer through the University of Sydney's student newsletter and the Centre of the Research and Innovation website[5]. An example of the flyer is presented in Appendix B.

All participants fulfilled both parts of the experiment (writing session and IAPS session). They were asked to come back one week after the experiment for s retrospective self-reporting for the writing session. One subject did not come back to complete the annotation and data for this participant was ignored.

6.3 Analysis of self-reported engagement and affect

6.3.1 Concurrent self-report (Engagement)

In the concurrent self-reports, the participants were asked to report their level of engagement verbally every two minutes in each writing session. As each session lasted an hour, 30 notifications were produced for each participant. Among 22 participants, two participants did not pay attention to any of them and did not report their level of engagement during the session. Other participants forgot to report their engagement level in some cases. In total, 530

[5] http://sydney.edu.au/research/involved/volunteer.shtml

responses were obtained in response to 660 notifications. Participants indicated that they were engaged for 425 cases (80%) compared with not being engaged (105 cases or 20%).

6.3.2 Retrospective self-report

Extracted video segments from the recorded videos in writing sessions were used for the retrospective self-reporting. Based on the procedure explained in Section 3-5-2, a researcher did the segmentation process. In total, 1325 video segments were extracted from 1320 minutes recording (22 participants × 60 minutes).

6.3.2.1 *Engagement*

For each participant on average, 60.23 (SD = 8.25) video segments were extracted. The average length of each video segment was 9.78 seconds (SD = 2.23). According to the participants' self-reports, they were engaged for a majority of the segments (996 instances or 75%). However, they reported not being engaged for 315 segments (24%). Fourteen instances (1%) were labelled as "Not Applicable (N/A)."

6.3.2.2 *Valence/Arousal*

Valence and arousal scores were reported in a 1–9 rating scale. The distribution of each reported score is presented in Figure 6-1. Almost normal distributions can be observed for reported valence (53%) and arousal (26%). The most frequent score was "score 5" for both dimensions (53% of valence and 26% of arousal scores).

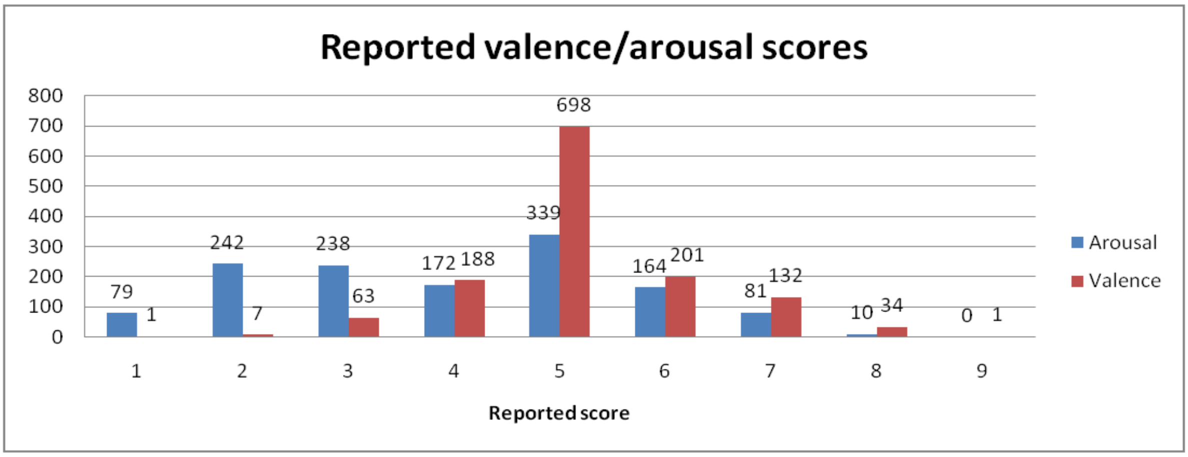


Figure 6-1: Distribution of reported scores in retrospective self-reports during writing sessions

Figure 6-1 shows a skewed class distribution towards the middle of the rating scale. To make a balanced distribution of classes ideal for creating a prediction model, the following calculations were conducted. First, the reported scores were standardized (converted to z-scores) for each participant. Then, reported values higher than the individual participant's mean were considered as "High Arousal" (or "Positive Valence") and the values smaller than or equal to participant's mean were considered as "Low Arousal" (or "Negative Valence"). Accordingly, the classification task will discriminate between two levels in valence and arousal dimension. Figure 6-2 shows the percentage of reported values in each group.

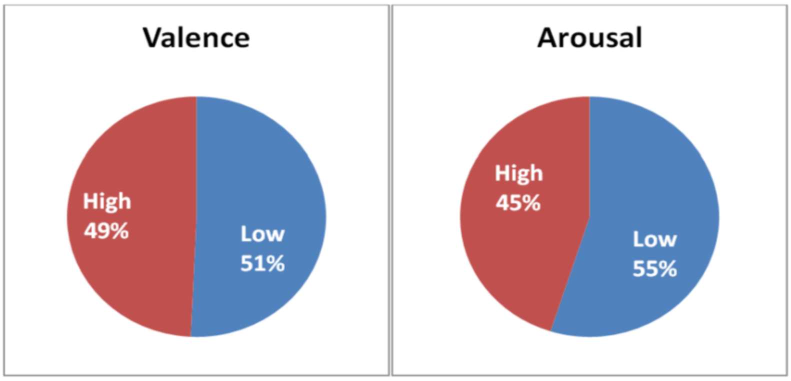


Figure 6-2: The percentage of reported valence and arousal after standardization and grouping into two levels (writing experiment)

6.3.2.3 *Categories*

Participants reported 1325 emotional categories retrospectively. Figure 6-3 presents the percentage of reporting each category using a pie chart (Calm 26%; Relaxed 23%; Bored 18%; Annoyed 5%; Glad 4%; Content 4%; Delighted 2%; Excited 2%; Depressed 1%; Gloomy 1%; Afraid 0%; Angry 0%; Others 13%). Calm and Relaxed, which could be considered as positive emotions, were reported more frequently in the writing experiment.

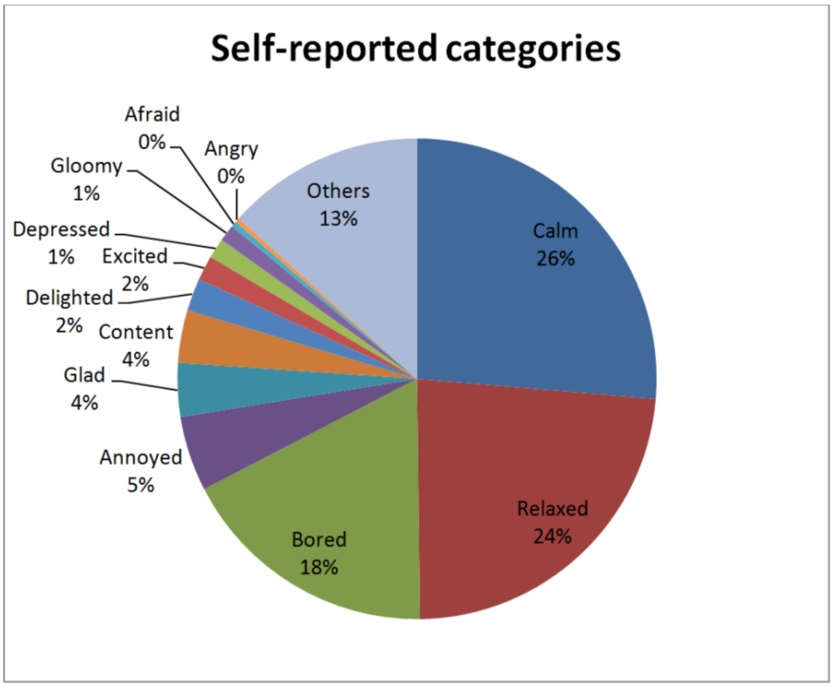


Figure 6-3: The percentage of each reported emotional category in the retrospective self-report during writing sessions

6.3.3 Engagement and affect during the writing session

To evaluate the reliability of annotations, we plotted the average engagement level of participants obtained from concurrent and retrospective self-reports over the time. A strong relationship between concurrent and retrospective engagement reports (the Pearson correlation was 0.82, $p < 0.001$) was observed, as shown in Figure 6-4. This provides a support for the reliability of the self-report measures.

According to Figure 6-4, on average 90% of participants were engaged at the beginning of the task and as they approached the middle of the task, their engagement levels decreased gradually. At the middle of the session, they had to submit their manuscripts and wait for 10 minutes to receive feedback. This decrease continued until t = 34 minutes and immediately after that time, their engagement level started to increase. The graph shows a peak at t = 42 minutes for concurrent engagement and again, their engagement waned as they neared the end of the session.

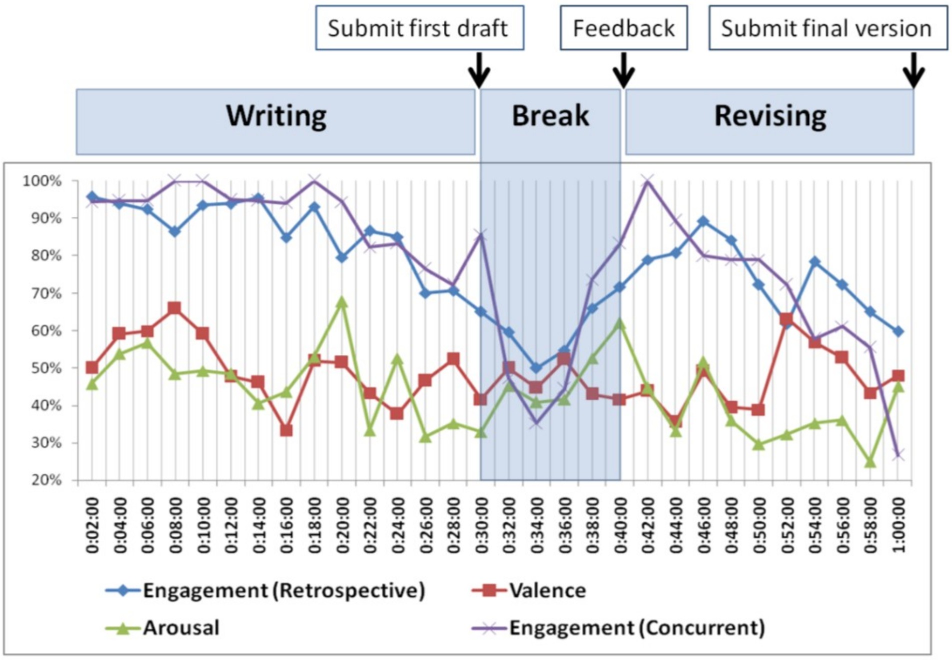


Figure 6-4: Average engagement level (from retrospective and concurrent reports) of participants and their emotions (valence/arousal) over the time

In Figure 6-4, the average changes of valence and arousal were plotted according to retrospective self-reports. This graph is based on two levels of valence and arousal dimensions. However, the graph shows random fluctuations of valence and arousal levels during the session. No significant correlation between affect (valence/arousal) and engagement was observed.

6.3.4 Impact of events on engagement and affect

We also analysed the impact of two types of interruptions during the writing session: feedback and distractions (loud beeps). To simulate some common distractions in an office, a loud beep was produced twice during the writing session. The amounts of reported engaged states before and after receiving the feedbacks are compared in Table 6-1. A significant difference between the engagement levels before (82%) and after (64%) receiving the feedback was observed as specified by a paired-samples *t* test (t(21) = 3.36, p < 0.005). However, the difference in participants' affect (valence and arousal) was not statistically significant before and after receiving feedback (valence: t(21) = 0.99, p = 0.33; arousal: t(21) = 1.43, p = 0.17).

We explored the impact of noises by analysing two minutes before and after each distraction. The results indicates that the engagement level was significantly reduced from 90% to 80% after each distraction (t(21) = 2.08 , p < 0.05). There was no significant differences for valence and arousal before and after distractions as specified by a paired-samples *t* test (valence: t(21) = 0.29, p = 0.77; arousal t(21) = –0.13, p = 0.89).

Table 6.1: Average values of reported engagement, valence and arousal before and after feedback and distractions

Self-reports	*Before feedback*	*After feedback*	*Before distraction (2 min)*	*After distraction (2 min)*
Engaged (%)	**82***	**64***	**90****	**80****
Not engaged (%)	**18**	**36**	**11**	**24**
Valence (1–9 average)	**5.18**	**5.08**	**5.14**	**5.11**
Arousal (1–9 average)	**4.03**	**3.85**	**4.26**	**4.28**

* *p < 0.005 : statistically significant*

** *p < 0.05 : statistically significant*

6.3.5 Affect in engaged vs. not-engaged states

In this section, we explore the relation between affective states and engagement level. The percentage of reported affective categories in each engagement level is presented in Table 6-2. Accordingly, "Bored" was the most reported affective state (49%) when participants were not engaged, whilst "Calm" was the most frequent affective state (29%) when engagement was reported by them.

Table 6.2: Descriptive statistics of reported emotion categories during engaged and not-engaged status

	Most reported categories	*Valence*	*Arousal*

Engaged	Calm (29%), Relaxed (25%), Others (15%), Bored (10%)	Mean = 5.22
		SD = 0.68
			**Mean = 4.01***
			SD = 1.34

Not-engaged	Bored (49%), Calm (18%), Relaxed (19%), Others (8%)	Mean =5.11
		SD = 0.78
			**Mean = 3.34***
			SD = 1.33

* *p < 0.005: statistically significant*

Although there was a slight difference for reported *valence* between engaged and not-engaged reports, a paired-samples *t* test showed that this difference was not significant (t(19) = 0.85, *p* = 0.41).

On the other hand, a paired-samples *t* test showed a significant difference of reported *arousal* scores between engaged reports (M = 4.01, SD = 1.34) compared with not-engaged reports (M = 3.34, SD = 1.33), t(19) = 4.77, *p* < 0.005). This finding can support this hypothesis: *People are more likely to arouse emotionally when they are engaged on a task.* Further analysis showed that the correlation between engagement and both valence (r = 0.13, *p* = 0.56) and arousal (r = 0.18, *p* = 0.44) were weak and non-significant.

6.4 Classification results

In this section, the performance of the vote classifier for discriminating two levels of valence, arousal and engagement, using different channels is reported. According to the methods explained in Chapter 3, for each video segment, 84 features were extracted using Microsoft Kinect Face Tracker (*FT*), 2,304 features were extracted using the LBPTOP method, and seven features were extracted for *HR* channel. Besides these three channels, two combinations of channels are also explored in this section. The first fusion model (*Fusion2*) combined HR and LBP-TOP channels. In the second fusion model (*Fusion3*), all three channels were combined.

Before analysing the extracted features, the features were synchronized with corresponding retrospective and concurrent labels. Prior to each classification task, the feature selection technique (CFS) was applied on the extracted features to reduce the dimensionality of the feature space. The results are reported separately for *retrospective* and *concurrent* labels. It should be mentioned that retrospective self-reports had labels for engagement, valence, and arousal compared with concurrent reports that only had engagement labels. First, the classification results for retrospective labels (valence, arousal, and engagement) are presented in two sub-sections: user-dependent and user-independent models. Then, in Section 6-4-2, the results for classifying concurrent reported engagement are also presented in user-dependent and user-independent sub-sections.

6.4.1 Retrospective affective states and engagement

Extracted labels from the retrospective self-reports were used for classification in this section. The participants rated each video segment in three aspects: valence, arousal, and engagement levels. The nine rating scales for valence and arousal were converted to the binary levels as described in Section 6-3-2-2. It should be mentioned that one participant labelled all instances with the same scale (e.g., 5) and there were not any instances in the "Positive Valence" group. For the engagement labels, two participants labelled all video instances with *Engaged* label. Those data sets could not provide valid data for classification. Those participants were ignored for classification. On the other hand, the LBPTOP method was not able to extract features from 75% of instances for two participants due to their appearances in front of the camera. One of the participants wore glasses with a thick frame and the eye-related features could not be extracted. The head position of another participant skewed toward the left and the face tracker could not track the face. Therefore, the classification results are reported based on recorded data from 17 participants.

6.4.1.1 *User-dependent models*

In the user-dependent analysis, 17 specific models were trained and tested for each participant. Table 6-3 reports the Kappa measures for each individual model to classify engagement, arousal, and valence levels reported in the retrospective self-reports. Figure 6-5 also presents the average Kappa measures for classifying affective states using three separate channels (HR, FT, LBP) along with the fusion models. The performance of each channel in detecting the two dimensions of affective states and engagement are discussed in this sub-section.

Table 6.3: The mean, standard deviation, maximum, and minimum of Kappa measures for classifying engagement, arousal (SelfAro), and valence (SelfVal) using three individual channels (HR, FT, and LBPTOP) and two fusion models (Fusion2, Fusion3)

	Engagement					SelfAro					SelfVal				
	HR	**FT**	**LBP**	**Fusion2**	**Fusion3**	**HR**	**FT**	**LBP**	**Fusion2**	**Fusion3**	**HR**	**FT**	**LBP**	**Fusion2**	**Fusion3**
Mean	**0.13**	**0.28**	**0.40**	**0.43**	**0.55**	**0.04**	**0.06**	**0.39**	**0.39**	**0.40**	**0.03**	**0.07**	**0.34**	**0.35**	**0.37**
Std	**0.24**	**0.29**	**0.39**	**0.38**	**0.31**	**0.23**	**0.24**	**0.23**	**0.23**	**0.25**	**0.18**	**0.16**	**0.24**	**0.23**	**0.24**
Max	**0.72**	**0.74**	**0.94**	**0.94**	**0.94**	**0.72**	**0.62**	**0.81**	**0.84**	**0.84**	**0.66**	**0.52**	**0.88**	**0.91**	**0.91**
Min	**-0.10**	**-0.03**	**-0.05**	**-0.05**	**0.00**	**-0.31**	**-0.42**	**0.00**	**0.00**	**0.00**	**-0.18**	**-0.09**	**0.00**	**0.00**	**-0.03**

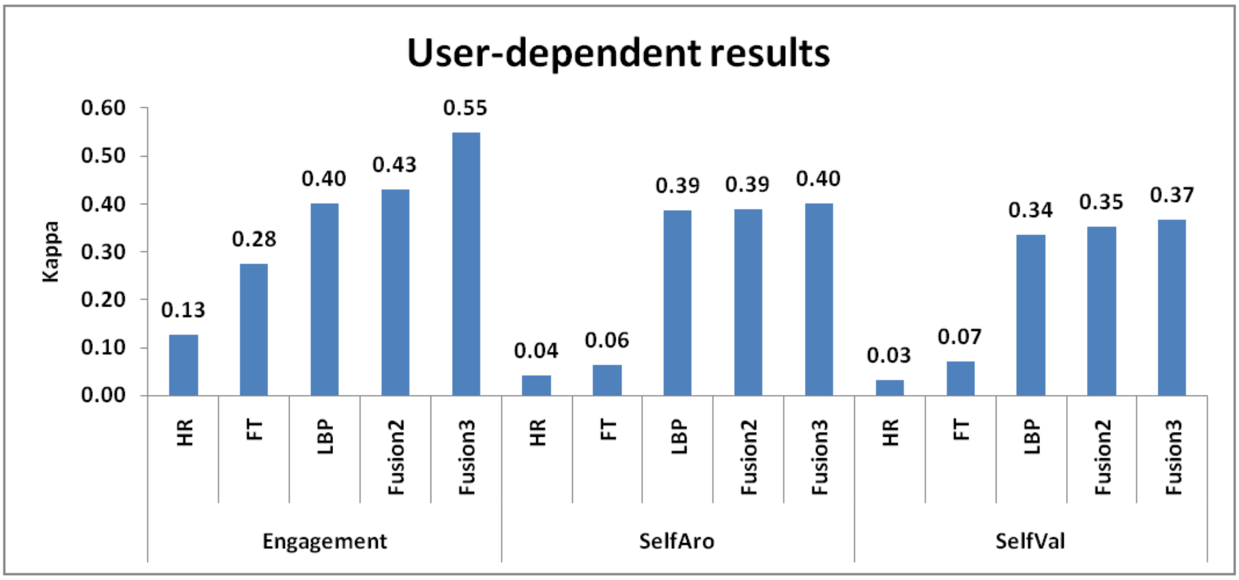


Figure 6-5: The average Kappa measures for detecting engagement, arousal, and valence (user-dependent models)

<u>**HR:**</u> On average, the HR channel obtained positive values of Kappa measure for detecting engagement, valence, and arousal. The best performance of this channel was for detecting engagement (Kappa = 0.13, SD = 0.24) across all participants. It should be mentioned that this channel achieved the maximum Kappa measure of 0.72 for one participant for engagement detection. The HR channel did not show a fair performance in arousal (Kappa = 0.04 SD = 0.23) and valence (Kappa = 0.03 SD = 0.18) detection.

<u>**FT:**</u> The Face Tracker channel also showed a fair accuracy in engagement detection with an average Kappa measure of 0.28 (SD = 0.29), whilst the results for arousal and valence detection were not promising. An average Kappa measure of 0.06 and 0.07 was achieved for classifying valence and arousal, respectively.

<u>**LBP:**</u> The LBPTOP channel was the most successful single channel among the three single channels for detecting affective states and engagement. This could demonstrate the impact of considering the dynamic features of facial expression in affect detection. The average Kappa measures of 0.40, 0.39, and 0.34 were obtained for detecting engagement, arousal, and

valence, respectively, which was a good performance. The paired *t* tests also specified that this channel significantly outperformed the HR channel ($p < 0.05$). The Kappa measures of LBP for detecting arousal and valence were significantly higher than the values achieved by the FT channel ($p < 0.05$). However, the paired two samples *t* test showed that the average Kappa measure achieved by LBP channel for engagement detection was not statistically higher than the Kappa measure obtained by the FT channel (t(16) = –1.30, $p = 0.21$).

Nevertheless, adding these two channels (HR and FT) to the LBP channel might increase the accuracy of affect detection. We explored the impact of adding HR and FT features to the LBP features by introducing two fusion models (*Fusion2*, *Fusion3*).

Fusion2: This fusion model combined LBP and HR features in the feature level. The feature selection function and the classification task were applied after this combination. The main reason for exploring this fusion model was to investigate the impact of adding HR features in affect detection. As specified by two samples *t* tests, the observed improvement of the *Fusion2* model over the LBP channel was not statistically significant (Engagement: t(16) = –1.66, $p = 0.12$; Arousal: t(16) = –1.00, $p = 0.33$; Valence: t(16) = –1.04, $p = 0.31$). These results suggested that adding the HR channel to the LBP channel would not improve the accuracy of affect detection.

Fusion3: This fusion model was created by adding the FT channel to other two channels used for the *Fusion2* model. For classifying engagement, the *Fusion3* model improved the average Kappa measure (achieved by *Fusion2* model) by 0.12, which was statistically significant as indicated by a paired two sample *t* test (t(16) = –2.51, $p < 0.05$). The Kappa measure of 0.55 obtained by the *Fusion3* model had the best accuracy among all other channels, and indicated a good agreement between predicted engagement levels and self-reported levels. However, the *Fusion3* model was not successful in improving the accuracy of arousal and valence detection compared with the *Fusion2* model, as indicated by paired two samples *t* tests (Arousal: t(16) = –0.96, *p* = 0.35; Valence: t(16) = –0.66, *p* = 0.52).

6.4.1.2 *User-independent model*

Building and analysing a general model for affect detection is always an important part of an automatic affect detection research. Here, the results of the user-independent models are reported. All instances from 17 participants used in the user-dependent analysis were standardized and combined to build the general model. The final model contained 1033 instances. The leave-one-out approach was used to validate the performance of our system. In each evaluation task, data from one participant were removed from the training set and used as a test set. The average Kappa measure for classifying affect and engagement are presented in Figure 6-6.

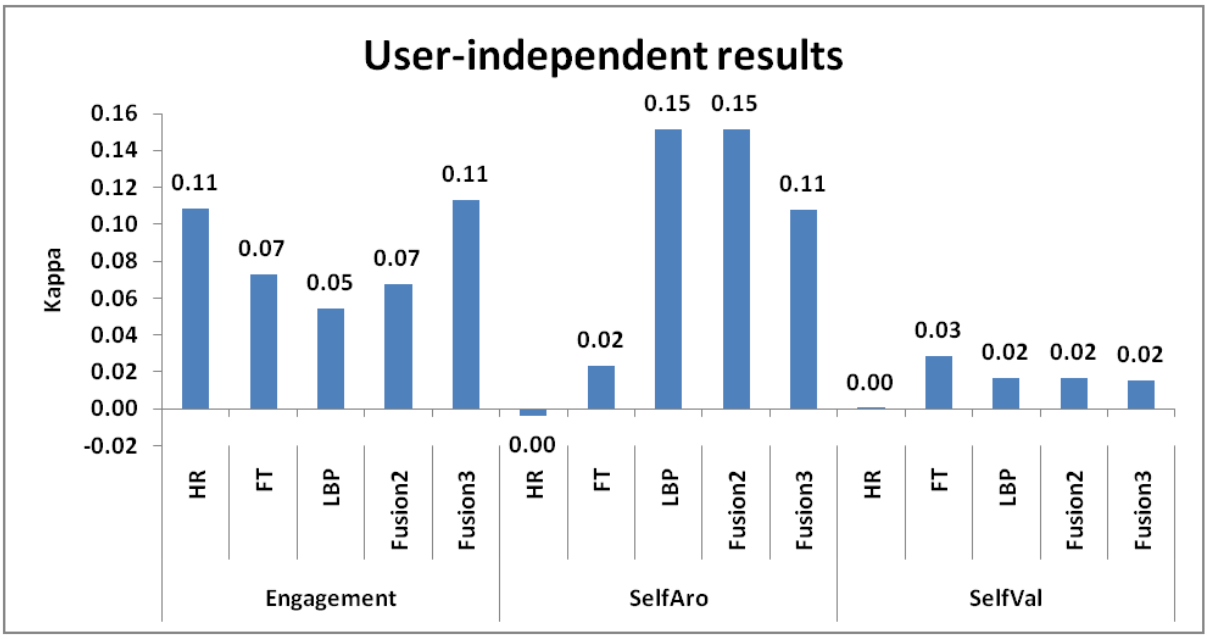


Figure 6-6: The Kappa measures for classifying affective states using leave-one-out cross validation approach (User-independent model)

The interesting result for the user-independent model was the performance of the HR channel for detecting engagement. The HR channel achieved the highest Kappa measure of 0.11.

Even the fusion models could not obtain better values for Kappa measure. As was suggested in the user-independent analysis (Section 6-4-1-1), the HR channel could be a good indicator for engagement detection. For arousal detection, the LBP channel and the *Fusion2* model obtained the best results (Kappa = 0.15). However, the reported Kappa measures for detecting valence were not promising for all single channels and fusion models.

6.4.2 Concurrent-reported engagement

This section reports the performance of the system for classifying the engagement levels reported concurrently during the writing sessions. As mentioned in Section 6-3-1, two participants forgot to report their level of engagement during the session and they were ignored in the feature extraction and classification. We extracted three types of features (HR, FT, and LBPTOP) for the rest of participants (19 participants). The user-dependent and user-independent models were built using these extracted data and are reported in following sub-sections.

Concurrent labels have been reported in specific times during the writing sessions. Before extracting the features, we needed to specify a time span for each label. We considered a 10-second window before each reported time and extracted all three types of features from those time windows.

6.4.2.1 *User-dependent models*

For each individual participant, a separate model was trained and validated using 10-fold cross validation approach. The average Kappa measures for discriminating *Engaged* and *Not-engaged* states were reported in Figure 6-7. Similar to previous analysis, besides single channels, two fusion models were also utilized for detecting engagement.

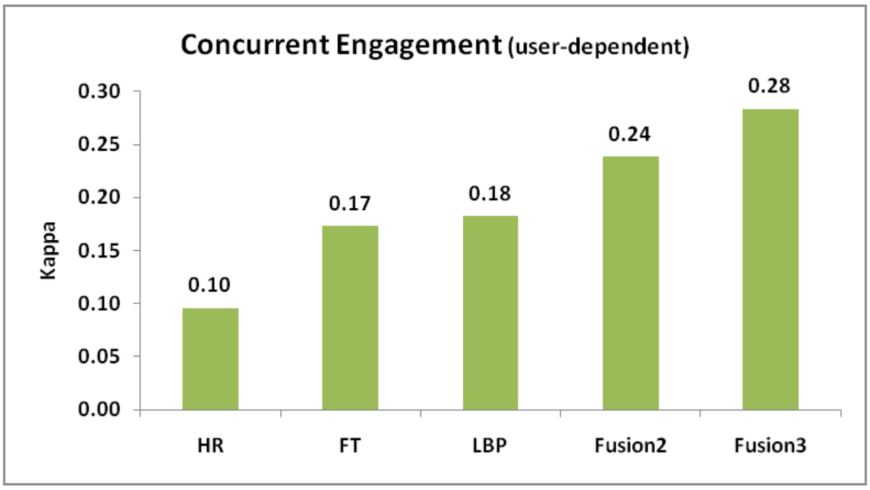


Figure 6-7: The average Kappa measure for classifying engagement levels reported concurrently (user-dependent models)

According to the average Kappa measures, combining three channels achieved the best performance (Kappa = 0.28) for detecting concurrent engagement. The improvement of the *Fusion3* model over the *HR* and *FT* channels were statistically significant ($p < 0.05$). However, paired two samples t tests indicated that the *Fusion3* model did not significantly improve the LBPTOP channel (t(18) = 1.66, p = 0.11) and the *Fusion2* model (t(18) = 1.39, p = 0.18).

The instances used in the concurrent self-reports were segmented automatically. This means the video segment might contain an incomplete part of a meaningful action or facial expression. Therefore, as reflected in the results, the obtained Kappa measures for detecting engagement were lower than the Kappa measures achieved by the retrospective reported engagement.

6.4.2.2 *User-independent model*

The extracted features from 19 participants were combined to build a general model for detecting concurrent reported engagement. Altogether, a data set with 509 instances was created. A leave-one-out approach was followed to evaluate the system and calculate the performance metrics. The Kappa measures obtained by each channel and two fusion models are reported in Figure 6-8.

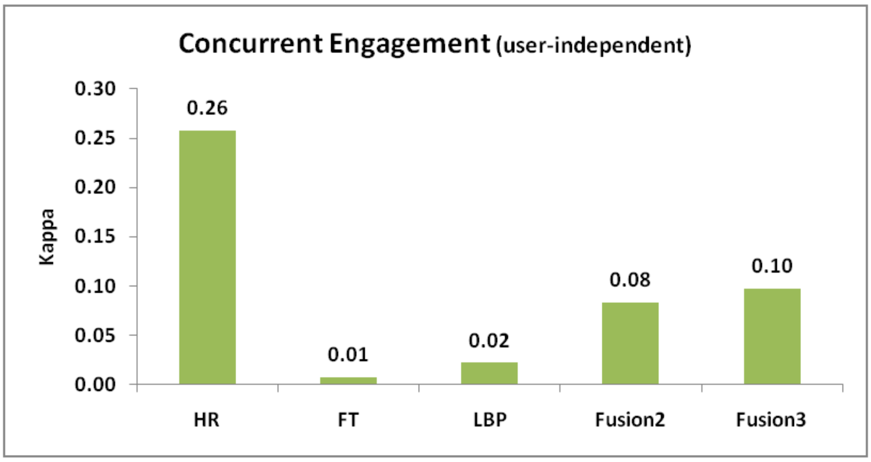


Figure 6-8: The Kappa measures for classifying concurrent reported engagement using leave-one-out cross validation approach (user-independent model)

Surprisingly, the HR channel showed the best performance. The Kappa measure of 0.26 was achieved by this channel was significantly higher than the Kappa measures obtained by the other two facial related channels: *FT* and *LBPTOP* ($p < 0.05$). This result shows the importance of the HR related features for detecting engagement, particularly when fixed segmentation has been used. The paired two samples *t* tests showed that the HR channel did not significantly overcome the *Fusion2* (t(18) = 2.02, $p = 0.06$) and *Fusion3* models (t(18) = 1.63, $p = 0.12$). According to Figure 6-8, adding the FT and LBPTOP channels to the HR features had a negative effect for detecting engagement.

6.5 Feature selection results

Analysing the features selected by the CFS gives us a better insight into important features for detecting valence, arousal, and engagement. This analysis shows the relation between different types of features from each channel (HR, FT, LBPTOP) and the participant's engagement and affective state. Finding the most important features can be useful to build the general model for affect detection. In the following subsections, we discuss the selected features for detecting affective states and engagement levels reported retrospectively and the engagement levels reported concurrently.

6.5.1 Retrospective affective states and engagement

In this section, we discuss the selected features for classifying retrospective self-reports as shown in Table 6-4. These features were selected from the *Fusion3* model, which contained all features from three channels. The selected features were grouped in three main categories based on the channels. According to the Table 6-4, the FT and LBPTOP features contributed to detecting engagement, arousal, and valence. However, the HR features were only selected for engagement detection. Figure 6-9 presents the distribution of appearance-based and motion-based features in the selected LBPTOP features. Figure 6-10 also shows the contribution of LBPTOP features based on the facial components: eyes and mouth.

Table 6.4: List of selected features by the CFS method for detecting retrospective reported engagement, valence and arousal (user-independent model)

	Selected features from Fusion model		
	Engagement	**SelfAro**	**SelfVal**
FT Features	ANU3-max, ANU4-max, Rx-mean, Rx-max, Ry-range, Rz-median, Tx-median, Tx-min, Tz-std, Tz-min	ANU3-diff, ANU4-max, Rx-median, Ty-max	ANU0-min, ANU0-diff, ANU1-std, ANU1-diff, ANU2-mean, ANU2-std, ANU3-diff, ANU5-median, Rx-mean, Ry-min, Tx-median, Tz-range
HR Features	HR-std, HR-min, HR-max, HR-range	–	–
LBPTOP Features	P35, P62, P87, P102, P109, P124, P150, P152, P160, P175, P197, P214, P221, P227, P737, P780, P786, P811, P814, P848, P858, P860, P862, P924, P926, P940, P941, P942, P949, P964, P996, P1013, P1016, P1085, P1445, P1515, P1609, P1611, P1625, P1638, P1646, P1654, P1655, P1709, P1715, P1726, P1746,	P26, P28, P74, P106, P120, P152, P182, P187, P227, P824, P828, P834, P853, P854, P860, P874, P875, P886, P939, P1457, P1559, P1628, P1646, P1701, P1747, P1953, P2034, P2059, P2123,	P74, P221, P234, P371, P806, P807, P810, P811, P812, P844, P861, P874, P886, P913, P931, P939, P941, P949, P981, P1003, P1025, P1571, P1622, P1646, P1699, P1747, P1807,

HR and FT: The selected HR features showed that changes in the HR in each video segment had a relationship with the engagement level. A combination of FT features contributed to engagement detection, which contained two features from animation units, four features from head rotation, and four features from the head translation. However, the distribution of selected ANU related features for valence detection was more than arousal detection. Almost all types of ANU features (except ANU4) were contributed in valence detection.

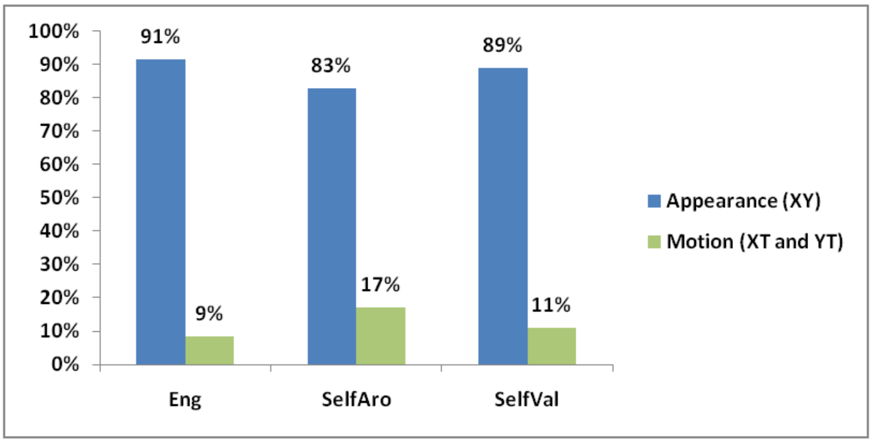


Figure 6-9: The percentage of contributions of appearance-based and motion-based LBPTOP features for engagement and affect detection (retrospective self-reports)

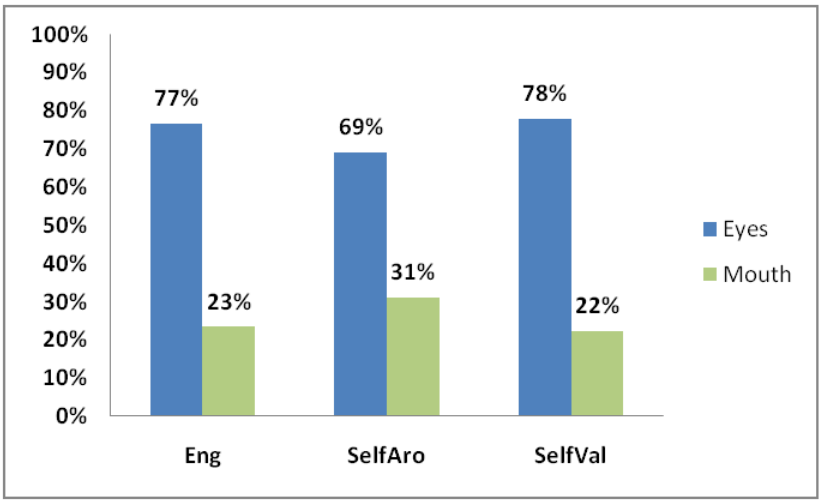


Figure 6-10: The percentage of contributions of the facial objects (eyes and mouth) in selected LBPTOP features for engagement and affect detection (retrospective self-reports)

LBPTOP: On the other hand, for engagement detection, most of the selected features among the LBPTOP features were from the appearance-based features. Eighty-four percent and 89% of selected features were from appearance-based features for detecting arousal and valence, respectively. According to Figure 6-9, motion-related features did not contribute much in engagement and affect detection. This might be because of the nature of the experiment, which was an HCI activity. Figure 6-10 shows that the eyes-related features were selected more for engagement detection compared with the mouth-related features. This was also true for arousal and valence detection. In general, exploring the selected features suggested that eyes-related and appearance-based features were good indicators for detecting engagement, valence, and arousal during writing sessions.

6.5.2 Concurrent engagement

The list of selected features for detecting concurrent reported engagement is reported in Table 6-5. Four HR features among seven features participated in the engagement detection. The *HR-std* and *HR-range* were also selected for detecting engagement, which were reported retrospectively. It suggests that measuring the changes of the HR can be a good indicator for engagement detection.

Table 6.5: List of selected features by the CFS method for detecting concurrent reported engagement (user-independent model)

	Selected features from Fusion model
	Concurrent Engagement
FT Features	ANU2-range, ANU5-std, Rx-max, Ty-range, Tz-min
HR Features	HR-mean, HR-median, HR-std, HR-range
LBPTOP Features	P20, P51, P92, P111, P118, P123, P150, P179, P199, P214, P245, P569, P778, P782, P844, P853, P1561, P1598, P1659, P1707, P1709, P1710, P1751,

The proportion of selected features from each type of LBPTOP features are presented in Figure 6-12. According to Figure 6-12, eyes-related features contributed more than mouth-related features for engagement detection, which was the same for retrospective reported engagement. In addition, the same result was observed for the appearance-based and motion-based features. Similar to the retrospective self-reports, appearance-based (96%) features were selected more than motion-based (4%) features for engagement detection.

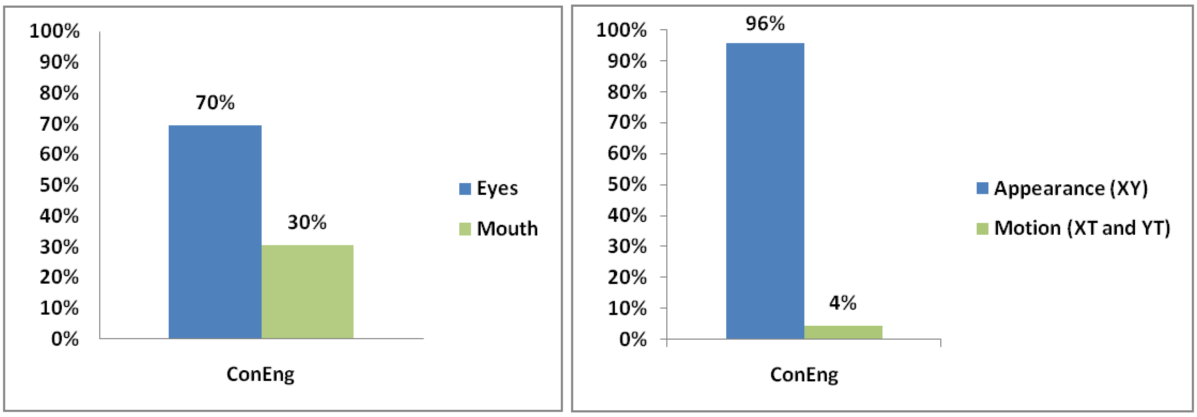


Figure 6-11: (a) The percentage of contributions of the facial objects (eyes and mouth) (b) The percentage of contributions of appearance-based and motion-based features in selected LBPTOP features for engagement detection (concurrent self-reports)

6.6 Conclusion

It has been shown that detecting engagement as a complex mental state and affective states is feasible using a multichannel approach. The results also showed that using the video-based HR measurement for affect detection is possible and that extracted features from this method can improve the accuracy of engagement detection. All of this information was recorded using a single Microsoft Kinect camera that can record video and depth at the same time. These kinds of sensors tend to be popular in the feature, and this method can be used in broad types of applications.

No significant correlation between engagement and affect was observed in our study. However, as expected, this study showed that peripheral distraction has a negative impact on engagement levels of the user. Current technologies can also monitor the contextual parameters and detect different distraction and noises (e.g., phone call, acoustic noises, etc.). Automatic detection of distractions can provide useful information for AC applications.

The combination of appearance-based (LBPTOP), geometric-based (Kinect FaceTracker), and physiological cues (HR) yielded the best result for classifying engagement that has been reported retrospectively. That combination (*Fusion3*) also outperformed all other channels for detecting concurrently reported engagement in user-dependent analysis. However, the HR channel showed better performance in a user-independent analysis that might indicate the universality of changes in HR compared with facial features that are more subjective. This result could show the importance of the HR related features for detecting engagement, particularly when fixed segmentation has been used.

A comprehensive analysis on selected features by the Correlation-based Feature Selection (CFS) method was provided at the end of this chapter. In general, analysing the selected features suggested that eyes-related and appearance-based features were good indicators for detecting engagement, valence, and arousal during writing activities.

Chapter 7. Conclusion

Summary

The main outcomes and contributions of the thesis are described in this chapter. The limitation of our proposed method and some suggestions for future work to develop a general computational model for recognizing complex affective states are also presented.

7.1 Outcomes

This thesis focused on developing a multichannel video-based framework for detecting naturalistic and non-basic emotions. Besides extracting traditional video-based features, such as facial expressions and head gestures, we proposed and evaluated a method for extracting physiological signals using the video modality. The combination of facial expression, head posture, and HR has been used for affect detection that was rarely explored in the AC research.

The fusion models outperformed the individual channels for affect detection in the semi-natural scenario with controlled stimulus presentation (second study). The fusion model also showed a small improvement over the individual channels using the user-dependent models in the naturalistic scenario (writing), while in the user-independent models; the fusion of channels did not improve the accuracy of affect detection. This finding showed the impact of the spontaneousness of the data in the accuracy of affect detection. As indicated by a comprehensive survey over the current multimodal affect detection systems (D'Mello & Kory, 2012), the multimodal approach can achieve a better performance over individual modalities when applied on the acted data compared with the natural data.

The ability of each channel or modality for detecting each affective state was assessed in this thesis. Some modalities or channels play a more significant role in affect detection compared with others. Recognizing channels that are more important can help us to create an effective combination of channels by assigning meaningful weights to each channel. Appearance-based, geometric-based, and chromatic features extracted from facial expressions and head gestures along with the HR features were evaluated. Unlike the basic emotions, more complex emotions, which occur commonly during human-computer interactions are not easily distinguishable by static features.

In this thesis, dynamic texture-based features were the most successful channel for non-basic affect detection in both controlled and naturalistic scenarios. This evidence showed the importance of the dynamic (temporal) changes in the appearance of the face for recognising non-basic affective states. Current studies also showed that considering dynamic changes is essential for discriminating basic emotions and non-basic ones. Although the HR channel did not show a good performance when used individually, some slights improvements were obtained by combining the HR channel with facial expressions.

Discovering the relation between each specific feature and each affective state is always a goal of affective computing research. For instance, to detect a smile, it is obvious that the system needs to be focused on the mouth region. The involvement rate of each set of features for detecting valence, arousal, and engagement has been investigated. Our results suggested that eye region contains more information about the engagement compared with the mouth region. There was also higher correlation between affective states and eye-related features in the naturalistic interactions.

Despite the current advances in remote-sensing physiological signals, a practical study to evaluate the possibility of using these techniques in affective computing applications was unexplored. We evaluated one of the state of the art methods (Poh et al., 2011) for remote HR measurement in three conditions and found that some modification is needed for use in naturalistic human computer interaction. A machine learning approach for improving that method has been proposed in this thesis and used for affect detection. The results indicated the possibility and usefulness of video-based HR monitoring system for affect detection.

Surprisingly, the HR channel showed a comparable accuracy in engagement detection using user-independent models. This might indicate that the physiological features were less subjective than facial features and that global physiological patterns could be defined for complex mental states like engagement.

Several studies have demonstrated that the appearance and temporal structure of posed expressions is different from natural/spontaneous ones (Cohn & Schmidt 2004; Hoque & Picard 2011). To have a robust and accurate system for recognizing natural expression and emotions, the system needs to be trained with natural data. In this thesis, the affect prediction models were trained and tested with two different experiments that contained naturalistic interactions. In the first scenario, a set of controlled stimulus was used to trigger natural emotional reactions of the users. Secondly, in a more naturalistic scenario, we have evaluated our system to detect users' emotions when they were writing using a personal computer system. As was expected, the system performed better in the first scenario compared with the second one, where the intensity of the affective states was lower due to the context of the experiment (writing task). The range of changes in valence and arousal levels during a normal writing is not so wide, so detecting the differentiation between two or more states with slight differences is not an easy task. On the other hand, affect detection systems might achieve a better accuracy when applied on an emotionally enriched application, such as action computer games.

One of the main goals of the affect detection studies is building a general model that can be applied on new and unseen users to detect their emotions accurately. It is a challenging issue, and most of the proposed systems could not outperform user-dependent models. In this thesis, user-dependent and user-independent models were developed and evaluated for detecting naturalistic affective states. Compared with the user-dependent models, the combination of individual channels did not improve the accuracy of affect detection for the user-independent models. By dividing general models into gender-specific models, the performance of individual channels for affect detection was improved significantly. This evidence suggests new approaches for building universal models. For example, by integrating a gender-specific affect detection system with an automatic gender detection system, a general model with higher-performance can be implemented.

Considering the context of the application is an important factor for detecting affective states in practical environments, environmental conditions and peripheral interventions can have either positive or negative impact on the user's affective states. Phone calls, email alerts, and other vocal noises occur commonly during human-computer interactions. Our findings showed that the peripheral noises could significantly decrease the engagement level of the users during the writing sessions. Accordingly, monitoring contextual variables using digital devices like cameras and microphones can give us cues regarding the user's affective and mental states.

7.2 Limitations

The face is the main source of extracting valuable information for affective states in the FER systems. It is always hard to record a full frontal view of the face in practical applications. Head movement and face occlusion are the main challenges associated with the FER systems. In our case, due to hand over the face, eyes occlusion by glasses, and fast and rigid head movements, the Kinect face tracker could not track the face accurately for some participants. Currently, novel methods have been proposed that can be used to detect affect through partially occluded faces (Zhi et al., 2011).

The position of the head was another problem in our experiment. Making sure that the subject and the camera were in the correct positions is a challenging task in practical applications. For example, one of the participants tended to be too close to the monitor during the writing session, which was out of the camera's field of view.

Changes in light conditions can introduce problems for video-based techniques. For example, continuous changing of sunlight on a partly cloudy day can change the amount of reflected light from the face, which produces noise for methods based on skin colour changes for either measuring heart pulses or detecting affective states.

Some people might not like to be filmed in every moment. This is a latent issue associated with all video-based methods. There is a trade-off between privacy and performance in using pervasive technologies. For instance, nobody likes to be traced by browser and internet search engines, but when they can reduce the searching time and offer them exactly what they are looking for, they will let their activities be traced! This is true for the video-based affect detection systems. If the system can detect emotions accurately and react accordingly, this issue tends to be resolved.

7.3 Future work

As has been mentioned in the outcomes section (Section 7.1), developing a general model for detecting non-basic and naturalistic affective states still needs more improvements. Current automatic affect detection systems perform well when they were trained with individuals but the reported results for user-independent models were not so promising. Eliciting more natural data from large population could be a solution in increasing the generalizability of the affect detection systems that could be considered as a future work.

Video-based methods are the most popular means used for multiple purposes (e.g., communications, surveillance, security, gaming, etc.), and the quality of services can be improved by adding the affect detection ability. Combinations of different video-based channels, such as facial expression, posture, gesture, and recently, physiological signals, is a way to improve the accuracy of the system that does not need extra equipment. New cameras with the ability to record in-depth information can be used for affect detection in special conditions, for example, in low light environments. In addition, other physiological features that can be extracted using video-based methods, such as Inter Beat Intervals and respiration rates, can also be used for affect detection.

All of the mentioned information captured by video modality can be combined with other modalities, such as voice, text, physiology, and contextual information to improve the accuracy of affect detection.

Current studies showed that the emotional model has an impact on the accuracy of the affect detection. Most of the current systems aimed to assign a label to a video segment. Video segmentation needs to be done before the classification process. This issue can be an obstacle to adopting an automatic affect prediction system in a practical application. It is also hard to divide a video segment into four common phases (neutral, onset, apex, and offset) that has been used for basic emotion. For a naturalistic affect, there might not be a clear difference between these four phases or even between two subsequent affective states. Developing new methods for automatic segmentation and continuous prediction of affective states is one of the essential directions of future works.

Most of the current datasets of natural affect were recorded in laboratory settings. New powerful and portable devices like smart-phones provide more opportunities to record natural data. New datasets for video-based affect detection using electronic handheld devices (e.g., smart phones, tablets, etc.) should be developed in the future.

Bibliography

Afzal, S., & Robinson, P. (2009). Natural Affect Data - Collection & Annotation in a Learning Context. In *3rd International conference on Affective Computing and Intelligent Interaction, ACII 2009* (pp. 1–7).

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37–66. doi:10.1007/BF00153759

Ahlberg, J. (2001). *CANDIDE-3 -- an updated parameterized face*. Sweden.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), R1–39. doi:10.1088/0967-3334/28/3/R01

Alzoubi, O., D'Mello, S. K., & Calvo, R. A. (2012). Detecting Naturalistic Expressions of Nonbasic Affect using Physiological Signals. *IEEE Transaction on Affective Computing*, *3*(3), 298–310.

Alzoubi, O., Hussain, M. S., & Calvo, R. A. (2014). Affect-Aware Assistive Technologies. In B. O'Neil & A. Gillepsie (Eds.), *Assistive Technology for Cognition*. University of Glasgow.

Anderson, K., & McOwan, P. W. (2006). A real-time automated system for the recognition of human facial expressions. *IEEE Transactions on Systems, Man, and Cybernetics. Part B,*, *36*(1), 96–105.

Andreassi, J. L. (2007). *Psychophysiology: Human Behavior and Physiological Response* (p. 538). Lawrence Erlbaum.

Arnold, M. B. (1960). *Emotion and Personality* (p. 430). Columbia Univ. Press.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson., R. (2009). Emotion Sensors go to School. In *14th Conference on Artificial Intelligence in Education* (pp. 17–24). Brighton, UK.

Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., & Solomon, P. E. (2009). The Painful Face - Pain Expression Recognition Using Active Appearance Models. *Image and Vision Computing*, *27*(12), 1788–1796. doi:10.1016/j.imavis.2009.05.007

Asthana, A., Saragih, J., Wagner, M., & Goecke, R. (2009). Evaluating AAM fitting methods for facial expression recognition. In *3rd International Conference on Affective Computing and Intelligent Interaction*. IEEE. doi:10.1109/ACII.2009.5349489

Averill, J. R. (1980). A Constructivist View of Emotion. In *Emotion: Theory, Research and Experience* (pp. 305–339). Academic Press.

Bailenson, J., Pontikakis, E., Mauss, I., Gross, J., Jabon, M., Hutcherson, C., … John, O. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, *66*(5), 303–317. doi:10.1016/j.ijhcs.2007.10.011

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

Banse, R., & Scherer, K. R. (1996). Acoustic Profiles in Vocal Emotion Expression. *J. Personality Social Psychology*, *70*(3), 614–636.

Bänziger, T., & Scherer, K. R. (2010). Introducing the geneva multimodal emotion portrayal (gemep) corpus. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for Affective Computing: A Sourcebook* (pp. 271–294). Oxford, U.K.: Oxford University Press.

Barreto, A., Zhai, J., & Adjouadi, M. (2007). Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. In *Proceedings of the IEEE international conference on Human-computer interaction* (pp. 29–38). Springer.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (pp. 568–573). doi:10.1109/CVPR.2005.297

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). Fully Automatic Facial Action Recognition in Spontaneous Behavior. In *7th IEEE International Conference on Automatic Face and Gesture Recognition (FGR06)* (pp. 223–230). doi:10.1109/FGR.2006.55

Bixler, R., & D'Mello, S. (2013). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 225–233).

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310.

Bohus, D., & Horvitz, E. (2009). Models for Multiparty Engagement in Open-World Dialog Models for Multiparty Engagement. In *Proceedings of SIGDIAL 2009* (pp. 225–234).

Borda, M. (2011). *Fundamentals in Information Theory and Coding* (p. 516). Springer.

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144–152.

Bourel, F., Chibelushi, C. C., & Low, A. a. (2002). Robust facial expression recognition using a state-based model of spatially-localised facial dynamics. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition* (pp. 113–118). IEEE.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., … Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *6th International Conference on Multimodal interfaces* (pp. 205–211). PA, USA: ACM New York.

Calvo, R. A., & D'Mello, S. (2010). Affect Detection : An Interdisciplinary Review of Models , Methods , and Their Applications. *IEEE Transaction on Affective Computing*, *1*(1), 18–37.

Calvo, R. A., & D'Mello, S. (Eds.). (2011). *New Perspectives on Affect and Learning Technologies*. *Explorations in the Learning Sciences, Instructional Systems and Performance Technologies* (Vol. 3). New York: Springer.

Cannon, W. B. (1927). The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology*, *39*(1), 106–124.

Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, *11*(1), 157–92.

Caridakis, G., Karpouzis, K., & Kollias, S. (2008). User and context adaptive neural networks for emotion recognition. *Neurocomputing*, *71*(13-15), 2553–2562. doi:10.1016/j.neucom.2007.11.043

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. *Affect and Emotion in Human-Computer Interaction. LNCS*, *4868*, 92–103.

Chanel, G., Ansari-Asl, K., & Pun, T. (2007). Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (pp. 2662–2667). Montreal, Quebec, Canada: IEEE.

Chang, Y., Hu, C., Feris, R., & Turk, M. (2006). Manifold based analysis of facial expression. *Image and Vision Computing*, *24*(6), 605–614. doi:10.1016/j.imavis.2005.08.006

Cheang, P. Y. S., & Smith, P. R. (2003). An Overview of Non-contact Photoplethysmography. *Electronic Systems and Control Division Research*, 57–59.

Chen, K. M., Misra, D., Wang, H., Chuang, H. R., & Postow, E. (1986). An X-band microwave life-detection system. *IEEE Transactions on Biomedical Engineering*, (7), 697–701.

Chetverikov, D., & Renaud, P. (2005). A brief survey of dynamic texture description and recognition. In *Proc. Conf. Computer Recognition Systems* (pp. 17–26).

Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012a). *Handbook of Research in Student Engagement*. New York: Springer.

Christenson, S. L., Reschly, A. L., & Wylie, C. (Eds.). (2012b). *Handbook of Research in Student Engagement*. New York: Springer.

Chu, W.-S., Torre, F. D. La, & Cohn, J. F. (2013). Selective Transfer Machine for Personalized Facial Action Unit Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 3515–3522. doi:10.1109/CVPR.2013.451

Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. S. (2003). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, *91*(1-2), 160–187. doi:10.1016/S1077-3142(03)00081-X

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

Cohn, J., & Schmidt, K. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, *2*, 1–12.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*(3), 287–314.

Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, *19*(3), 267–303.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, *18*(1), 32–80.

Craig, S. D., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning. *Cognition & Emotion*, *22*(5), 777–788. doi:10.1080/02699930701516759

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Learning, Media & Technology*, *29*(3), 241–250. doi:10.1080/1358165042000283101

D'Mello, S., & Calvo, R. A. (2013). Beyond the Basic Emotions : What Should Affective Computing Compute ? In *CHI 2013 - Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris, France.

D'Mello, S., Chipman, P., & Graesser, A. (2007). Posture as a Predictor of Learner's Affective Engagement. In *Proceedings of the 29th Annual Cognitive Science Society* (Vol. 1, pp. 905–910).

D’Mello, S., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, *20*(2), 147–187. doi:10.1007/s11257-010-9074-4

D’Mello, S., Graesser, A., & Picard, R. W. (2007). Toward an Affect- Sensitive AutoTutor. *IEEE Intelligent Systems*, *22*(4), 53–61.

D’Mello, S., & Kory, J. (2012). Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *14th ACM International Conference on Multimodal Interaction* (pp. 31–38). Santa Monica, California, USA.

Dalgleish, T., Dunn, B., & Mobbs, D. (2009). Affective Neuroscience: Past, Present, and Future. *Emotion Review*, *1*, 355–368.

Damasio, A. R. (2008). *Descartes’ Error: Emotion, Reason, and the Human Brain.* (p. 352). Random House.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals* (p. 374). John Murray.

De Vicente, A. (2003). *Towards tutoring systems that detect students’ motivation: an investigation*. University of Edinburgh, UK.

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying Facial Actions. *IEEE Pattern Analysis and Machine Intelligence*, *21*(10), 974–989.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., & McRorie, M. (2007). The HUMAINE Database: addressing the needs of the affective computing community. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction* (pp. 488–500). Lisbon, Portugal.

Dziuda, Ł., Skibniewski, F. W., Krej, M., & Baran, P. M. (2013). Fiber Bragg grating-based sensor for monitoring respiration and heart activity during magnetic resonance imaging examinations. *Journal of Biomedical Optics*, *18*(5). doi:10.1117/1

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion*, *6*, 169–200.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124–9.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, 1978.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (2003). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues* (p. 212). ISHK.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial Action Coding System*. A Human Face.

Ekman, P., & Rosenberg, E. L. (2005). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)* (p. 672). Oxford University Press, USA.

El Kaliouby, R. (2005). *Mind-reading machines: automated inference of complex mental states*. University of Cambridge.

El Kaliouby, R., & Robinson, P. (2005). Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *Real-Time Vision for HCI* (pp. 181–200). Spring-Verlag. doi:10.1109/CVPR.2004.427

Elfenbein, H. A., & Ambady, N. (2003). Universals and cultural differences in recognizing emotions of a different cultural group. *Current Directions in Psychological Science.*, *12*(5), 159–164.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern Recognition*, *36*(1), 259–275. doi:10.1016/S0031-3203(02)00052-3

Fei, J., & Pavlidis, I. (2010). Thermistor at a distance: unobtrusive measurement of breathing. *IEEE Transactions on Bio-Medical Engineering*, *57*(4), 988–98. doi:10.1109/TBME.2009.2032415

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Forbes-riley, K., & Litman, D. (2011). When Does Disengagement Correlate with Learning in Spoken Dialog Computer Tutoring ? In *Proceedings of the 15th international conference on Artificial intelligence in education* (pp. 81–89).

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, *74*(1), 59–109. doi:10.3102/00346543074001059

Frijda, N. (1987). Emotion, Cognitive Structure, and Action Tendency. *Cognition and Emotion*, *1*(11), 115–143.

Garbey, M., Sun, N., Merla, A., & Pavlidis, I. (2004). Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. *IEEE Transactions on Bio-Medical Engineering*, *57*(4), 1418–1426.

Graesser, A., Witherspoon, A., McDaniel, B., D'Mello, S., Chipman, P., & Gholson, B. (2006). Detection of Emotions during Learning with AutoTutor. In *Proceedings of the 28th Annual Meetings of the Cognitive Science Society* (pp. 285–290).

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013). Automatically Recognizing Facial Expression : Predicting Engagement and Frustration. In *Proceedings of the 6th International Conference on Educational Data Mining*.

Greneker, E. F. (1997). Radar sensing of heartbeat and respiration at a distance with applications of the technology. In *RADAR* (pp. 150 – 154). Edinburgh, U.K.

Gunes, H., & Pantic, M. (2010). Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Intelligent Virtual Agents* (pp. 371–377). Springer.

Gunes, H., & Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, *31*(2), 120–136. doi:10.1016/j.imavis.2012.06.016

Guo, G., & Dyer, C. R. (2005). Learning from examples in the small sample case: face expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics. Part B,*, *35*(3), 477–88.

Hall, M. A. (2000). Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 359–366).

Hamming, R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, *29*(2), 147–160.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining* (p. 546). Cambridge, MA: MIT Press.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. doi:10.1109/TKDE.2008.239

Healey, J. a., & Picard, R. W. (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. doi:10.1109/TITS.2005.848368

Hertzman, A. B., & Spealman, C. R. (1937). Observations on the finger volume pulse recorded photo- electrically. *American Journal of Physiology*, *119*, 334–335.

Hoque, M. E., Kaliouby, R., & Picard, R. W. (2009). When Human Coders (and Machines) Disagree on the Meaning of Facial Affect in Spontaneous Videos. In *9th International Conference on Intelligent Virtual Agents, (IVA 2009)* (pp. 337–343). Amesterdam.

Hoque, M. E., McDuff, D. J., & Picard, R. W. (2012). Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles. *IEEE Transaction on Affective Computing*. doi:10.1109/T-AFFC.2012.11

Hoque, M., & Picard, R. W. (2011). Acted vs. natural frustration and delight: Many people smile in natural frustration. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)* (pp. 354–359).

Hummler, H. D., Engelmann, A., Pohlandt, F., Högel, J., & Franz, A. R. (2004). Accuracy of pulse oximetry readings in an animal model of low perfusion caused by emerging pneumonia and sepsis. *Intensive Care Medicine*, *30*(4), 709–13. doi:10.1007/s00134-003-2116-1

Hussain, M. S., & Calvo, R. A. (2011). Multimodal Affect Detection from Physiological and Facial Features during ITS Interaction. In *The 15th International Conference on Artificial Intelligence in Education (AIED)* (pp. 472–474). Auckland, New Zealand.

Hussain, M. S., Calvo, R. A., & Chen, F. (2013). Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion During Affective Interference. *Interacting with Computers*. doi:10.1093/iwc/iwt032

Hussain, M. S., Calvo, R. A., & Pour, P. A. (2011). Hybrid Fusion Approach for Detecting Affects from Multichannel Physiology. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective Computing and Intelligent Interaction* (pp. 568–577). Springer.

Hussain, M. S., Monkaresi, H., & Calvo, R. A. (2012a). Categorical vs. Dimensional Representations in Multimodal Affect Detection during Learning. In *11th International Conference on Intelligent Tutoring Systems (ITS '12)* (pp. 78–83). Berlin / Heidelberg: Springer. doi:10.1007/978-3-642-30950-2_11

Hussain, M. S., Monkaresi, H., & Calvo, R. A. (2012b). Combining Classifiers in Multimodal Affect Detection. In *Proceedings of the Tenth Australasian Data Mining Conference (AusDM 2012)* (pp. 103–108). Sydney, Australia.

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding*, *108*(1-2), 116–134. doi:10.1016/j.cviu.2006.10.019

James, W. (1884). What Is an Emotion? *Mind*, *9*, 188–205.

Joshi, M. V. (2002). On evaluating performance of classifiers for rare classes. In *IEEE International Conference on Data Mining* (pp. 641–644). IEEE Comput. Soc. doi:10.1109/ICDM.2002.1184018

Kahn, W. A. (1990a). Psychological Conditions of Personal Engagement and Disengagement at Work. *Academy of Management Journal*, *33*(4), 692–724.

Kahn, W. A. (1990b). Psychological Conditions of Personal Engagement and Disengagement at Work. *Academy of Management Journal*, *33*(4), 692–724.

Kahneman, D. (2011). *Thinking, fast and slow* (p. 512). New York: Farrar, Straus and Giroux.

Kanade, T., Cohn, J., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition* (pp. 46–53).

Killingsworth, M. A., & Gilbert, D. T. D. (2010). A wandering mind is an unhappy mind. *Science*, *330*(6006), 932. doi:10.1126/science.1192439

Kim, J. (2007). Bimodal emotion recognition using speech and physiological changes. In M. Grimm & K. Kroschel (Eds.), *Robust Speech Recognition and Understanding* (pp. 265–280). Vienna, Austria: I-Tech Education and Publishing.

Kim, J., & Andre, E. (2006). Emotion recognition using physiological and speech signal in short-term observation. In E. André, L. Dybkjær, W. Minker, H. Neumann, & M. Weber (Eds.), *Perception and Interactive Technologies. LNCS* (Vol. 4021, pp. 53–64). Springer-Berlin Heidelberg.

Koelstra, S., Pantic, M., & Patras, I. (2010). A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(11), 1940–54. doi:10.1109/TPAMI.2010.50

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

Kreibig, S. D. (2010). Autonomic Nervous System Activity in Emotion: A Review. *Biological Psychology*, *84*, 394–421.

Kuncheva, L. I. (2002). Switching between selection and fusion in combining classifiers: an experiment. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *32*(2), 146–56. doi:10.1109/3477.990871

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms* (p. 350). Wiley Inter-science.

Lang, P., & Bradley, M. (1997). International affective picture system (IAPS): Technical manual and affective ratings. *Psychology*.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual* (p. 61). Gainesville, FL.

Larsen, R. J. (2000). Target Articles: Toward a Science of Mood Regulation. *Psychological Inquiry*, *11*(3), 129–141.

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

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary Facial Action Generates Emotion-Specific Autonomic Nervous System Activity. *Psychophysiology*, *27*, 363–384.

Li, C., Cummings, J., & Lam, J. (2009). Radar remote monitoring of vital signs. *Microwave Magazine,*, *10*(February), 47–56.

Lichtenstein, A., Oehme, A., Kupschick, S., & Jürgensohn, T. (2008). Comparing two emotion models for deriving affective states from physiological data. In C. Peter & R. Beale (Eds.), *Affect and Emotion in Human-Computer Interaction* (pp. 35–50). Springer Berlin Heidelberg.

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, *24*(6), 615–625. doi:10.1016/j.imavis.2005.09.011

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)* (pp. 298–305). doi:10.1109/FG.2011.5771414

Liu, C., Conn, K., Sarkar, N., & Stone, W. (2008). Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder. *International Journal of Human-Computer Studies*, *66*(9), 662–677. doi:10.1016/j.ijhsc.2008.04.003

Liu, M., Calvo, R. A., & Pardo, A. (2013). Tracer : A tool to measure and visualize student engagment in writing activities. In *IEEE 13th International Conference onAdvanced Learning Technologies (ICALT)* (pp. 421–425).

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., & Ave, F. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 94–101).

Lucey, S., Ashraf, A. B., & Cohn, J. F. (2007). Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face. In K. Delac & M. Grgic (Eds.), *Face Recognition* (pp. 275–286). I-Tech Education and Publishing.

Mahmoud, M., Baltrusaitis, T., Robinson, P., & Riek, L. D. (2011). 3D Corpus of Spontaneous Complex Mental States. In S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective Computing and Intelligent Interaction* (pp. 205–214). Springer.

Mangai, U., Samanta, S., Das, S., & Chowdhury, P. (2010). A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification. *IETE Technical Review*, *27*(4), 293. doi:10.4103/0256-4602.64604

Mann, S. (1997). Wearable computing: A first step toward personal imaging. *Computer*, *30*(2), 25–32.

McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., & Graesser, A. (2007). Facial Features for Affective State Detection in Learning Environments. In *29th Annual meeting of the cognitive science society* (pp. 467–472). Austin, TX: Cognitive Science Society.

Mckeown, G., Valstar, M. F., Cowie, R., & Pantic, M. (2010). The SEMAINE corpus of emotionally coloured character interactions. In *IEEE International Conference on Multimedia and Expo* (pp. 1079–1084).

Mehrabian, A. (1968). Communication without words. *Psychol. Today*, *2*(4), 53–56.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Michel, P., & El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. *Proceedings of the 5th International Conference on Multimodal Interfaces - ICMI '03*, 258. doi:10.1145/958468.958479

Monkaresi, H., Calvo, R. A., & Hussain, M. S. (2012). Automatic natural expression recognition using head movement and skin color features. In Genny Tortora, Stefano Levialdi, & Maurizio Tucci (Eds.), *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)* (pp. 657–660). New York: ACM. doi:10.1145/2254556.2254678

Monkaresi, H., Calvo, R. A., & Hussain, M. S. (2014). Using Remote Heart Rate Measurement for Affect Detection. In *The 27th International FLAIRS Conference* (pp. 118–123). Pensacola Beach, Florida, USA: AAAI.

Monkaresi, H., Calvo, R. A., & Yan, H. (2014). A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam. *IEEE Journal of Biomedical and Health Informatics*, *18*(4), 1153–1160. doi:10.1109/JBHI.2013.2291900

Monkaresi, H., Hussain, M. S., & Calvo, R. A. (2012a). A Dynamic Approach for Detecting Naturalistic Affective States from Facial Videos during HCI. In M. Thielscher & D. Zhang (Eds.), *AI 2012: Advances in Artificial Intelligence* (pp. 170–181). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-35101-3_15

Monkaresi, H., Hussain, M. S., & Calvo, R. A. (2012b). Classification of Affects Using Head Movement, Skin Color Features and Physiological Signals. In *IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2012)* (pp. 2664–2669). doi:10.1109/ICSMC.2012.6378149

Moore, S., & Bowden, R. (2011). Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, *115*(4), 541–558. doi:10.1016/j.cviu.2010.12.001

Moridis, C. N., & Economides, A. a. (2009). Mood Recognition during Online Self-Assessment Tests. *IEEE Transactions on Learning Technologies*, *2*(1), 50–61. doi:10.1109/TLT.2009.12

Nakano, Y. I., & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '10* (pp. 139–148). New York, New York, USA: ACM Press. doi:10.1145/1719970.1719990

Nguyen, T., Bass, I., Li, M., & Sethi, I. (2005). Investigation of combining SVM and decision tree for emotion classification. In *The Seventh IEEE International Symposium on Multimedia (ISM'05)* (pp. 540–544).

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011). Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence–Arousal Space. *IEEE Transactions on Affective Computing*, *2*(2), 92–105.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, *29*(1), 51–59.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, *97*, 315–331.

Pal, P., Iyer, A. N., & Yantorno, R. E. (2006). Emotion Detection from Infant Facial Expressions and Cries. In *IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP '06)* (pp. 721–724).

Pantelopoulos, a., & Bourbakis, N. G. (2010). A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(1), 1–12. doi:10.1109/TSMCC.2009.2032660

Pantic, M., & Patras, I. (2006). Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments. *IEEE Transactions on Systems, Man, and Cybernetics*, *36*(2), 433–449.

Pantic, M., & Rothkrantz, L. J. M. (2000a). Automatic Analysis of Facial Expressions : The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1424–1445.

Pantic, M., & Rothkrantz, L. J. M. (2000b). Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, *18*(11), 881–905. doi:10.1016/S0262-8856(00)00034-2

Pantic, M., & Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1370–1390. doi:10.1109/JPROC.2003.817122

Pantic, M., Sebe, N., Cohn, J. F., & Huang, T. (2005). Affective multimodal human-computer interaction. *Proceedings of the 13th Annual ACM International Conference on Multimedia - MULTIMEDIA '05*, 669. doi:10.1145/1101149.1101299

Pantic, M., Valstar, M. F., Rademaker, R., & Maat, L. (2005). Web-Based Database for Facial Expression Analysis. In *IEEE International Conference on Multimedia and Expo* (pp. 317–321). doi:10.1109/ICME.2005.1521424

Peter, C., Ebert, E., & Beikirch, H. (2009). *Physiological Sensing for Affective Computing*. Springer London.

Peters, C., Castellano, G., & de Freitas, S. (2009). An exploration of user engagement in HCI. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots - AFFINE '09* (pp. 1–3). New York, New York, USA: ACM Press. doi:10.1145/1655260.1655269

Picard, R. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies*, *59*(1-2), 55–64. doi:10.1016/S1071-5819(03)00052-1

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

Picard, R. W. (2000). *Affective Computing* (p. 304). The MIT Press.

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., … Strohecker, C. (2004). Affective Learning — A Manifesto. *BT Technology Journal*, *22*(4), 253–269. doi:10.1023/B:BTTJ.0000047603.37042.33

Picard, R. W., Wexelblat, A., Clifford, I. N., & Clifford, I. N. I. (2002). Future interfaces: social and emotional. In *CHI'02 extended abstracts on Human factors in computing systems* (pp. 698–699). ACM.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. In B. Schoelkopf, C. Burges, & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning* (pp. 1–21).

Poh, M. Z., McDuff, D. J., & Picard, R. W. (2010). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, *18*(10), 10762–74.

Poh, M. Z., McDuff, D. J., & Picard, R. W. (2011). Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Bio-Medical Engineering*, *58*(1), 7–11. doi:10.1109/TBME.2010.2086456

Polikar, R. (2006). Pattern recognition. In M. Akay (Ed.), *Wiley Encyclopedia of Biomedical Engineering* (pp. 1–22). New York, USA: Wiley.

Porayska-pomsta, K., Mavrikis, M., D'Mello, S. K., Conati, C., & Baker, R. (2013). Knowledge Elicitation Methods for Affect Modelling in Education. *International Journal of Artificial Intelligence in Education*, *22*(3), 107–140.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning* (p. 302). Morgan kaufmann.

Robson, C. (2011). *Real Word Research: A Resource for Social Scientist and Practitioner Researchers* (3rd ed., p. 608). Wiley.

Roda, C., & Thomas, J. (2006). Attention aware systems: Theories, applications, and research agenda. *Computers in Human Behavior*, *22*(4), 557–587. doi:10.1016/j.chb.2005.12.005

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: theory and applications* (p. 244). World Scientific Pub Co Inc.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*. doi:10.1037/h0077714

Russell, J. A. (1991). Culture and the Categorization of Emotions. *Psychological Bull.*, *110*(3), 426–450.

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Rev.*, *110*, 145–172.

Saisan, P., Doretto, G., & Wu, Y. (2001). Dynamic texture recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (pp. 58–63).

Sanou, B. (2013). *The world in 2013: ICT Facts and Figures* (p. 8). Geneva, Switzerland.

Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., & Parrott, D. J. (2001). A psychometric evaluation of facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior*, *25*(3), 167–185.

Scherer, K. R., & Ekman, P. (1982). *Handbook of methods in nonverbal behavior research*. Cambridge, UK: Cambridge University Press.

Schuller, B., Muller, R., Hornler, B., Hothker, A., Konosu, H., & Rigoll, G. (2007). Audiovisual Recognition of Spontaneous Interest within Conversations. In *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07)* (pp. 30–37).

Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing*, *27*(6), 803–816. doi:10.1016/j.imavis.2008.08.005

Shin, Y. (2007). Facial Expression Recognition Based on Emotion. In Y. Shi, G. D. van Albada, J. Dongarra, & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2007* (pp. 81–88). Springer Berlin Heidelberg.

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing*, *3*(1), 42–55. doi:10.1109/T-AFFC.2011.25

Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing*, *3*(2), 211–223. doi:10.1109/T-AFFC.2011.37

Sprager, S., & Zazula, D. (2013). Detection of heartbeat and respiration from optical interferometric signal by using wavelet transform. *Computer Methods and Programs in Biomedicine*, *111*(1), 41–51. doi:10.1016/j.cmpb.2013.03.003

Takano, C., & Ohta, Y. (2007). Heart rate measurement based on a time-lapse image. *Medical Engineering & Physics*, *29*(8), 853–7. doi:10.1016/j.medengphy.2006.09.006

Tarvainen, M. P., Ranta-aho, P. O., & Karjalainen, P. A. (2002). An advanced detrending method with application to HRV analysis. *IEEE Trans. Biomed. Eng.*, *49*(2), 172–175.

Tian, Y., Kanade, T., & Cohn, J. F. (2001). Recognizing Action Units for Facial Expression Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *23*(2), 97–115.

Tian, Y., Kanade, T., & Cohn, J. F. (2002). Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition,* (pp. 229 – 234).

Tong, Y., Liao, W., & Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(10), 1683–99. doi:10.1109/TPAMI.2007.1094

Torre, F. De, Simon, T., Ambadar, Z., & Cohn, J. F. (2011). FAST-FACS : A Computer-Assisted System to Increase Speed and Reliability of Manual FACS Coding. In *4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011)* (pp. 57–66). Memphis, TN.

Trotter, H. F. (1959). An elementary proof of the central limit theorem. *Archiv Der Mathematik*, *10*(1), 226–234. doi:10.1007/BF01240790

Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., & Scherer, K. (2011). The first facial expression recognition and analysis challenge. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)* (pp. 921–926). IEEE. doi:10.1109/FG.2011.5771374

Valstar, M. F., Mehu, M., Jiang, B., Pantic, M., & Scherer, K. (2012). Meta-Analysis of the First Facial Expression Recognition Challenge. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics : A Publication of the IEEE Systems, Man, and Cybernetics Society*, *42*(4), 966–979. doi:10.1109/TSMCB.2012.2200675

Valstar, M. F., & Pantic, M. (2007). Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics. In *Lecture Notes on Computer Science* (Vol. 4796, pp. 118– 127). Springer.

Valstar, M. F., & Pantic, M. (2012). Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *42*(1), 28–43. doi:10.1109/TSMCB.2011.2163710

Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. *Optics Express*, *16*(26), 21434–45.

Vertegaal, R. (2003). Attentive User Interfaces. *Communications of the ACM*, *46*(3), 30–33.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '01)* (pp. 511–518).

Wagner, J., Kim, J., & André, E. (2005). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005.* (pp. 940–943).

Werbos, P. J. (1994). The Brain as a Neurocontroller: New Hypotheses and New Experimental Possibilities. In K. Pribram (Ed.), *Origins: Brain and Self-Organization* (pp. 680–706).

Whitehill, J., & Omlin, C. W. (2006). Haar Features for FACS AU Recognition. In *7th IEEE International Conference on Automatic Face and Gesture Recognition (FGR06)* (pp. 97–101). IEEE. doi:10.1109/FGR.2006.61

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (p. 525). Morgan Kaufmann.

Yu, C., Aoki, P., & Woodruff, A. (2004). Detecting user engagement in everyday conversations. In *8th International Conference on Spoken Language Processing (ICSLP 2004)* (pp. 1329–1332). Jeju Island; Korea.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58. doi:10.1109/TPAMI.2008.52

Zeng, Z., Tu, J., Liu, M., Huang, T. S., Pianfetti, B., Roth, D., & Levinson, S. (2007). Audio-Visual Affect Recognition. *IEEE Trans. Multimedia*, *9*(2), 424–428.

Zhang, Z., Lyons, M., Schuster, M., & Akamatsu, S. (1998). Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-lyaer perceptron. In *Proc. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition* (pp. 454–459).

Zhao, G., & Pietikäinen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine*, *29*(6), 915–928.

Zhao, G., & Pietikäinen, M. (2009). Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. *Pattern Recognition Letters*, *30*(12), 1117–1127. doi:10.1016/j.patrec.2009.03.018

Zhi, R., Flierl, M., Ruan, Q., & Kleijn, W. B. (2011). Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *41*(1), 38–52. doi:10.1109/TSMCB.2010.2044788

–

APPENDIX A

The list of IAPS pictures used for emotion elicitation in Chapter 5

(Male subjects)

Male subjects	Low Valence - Low Arousal				
		Valence		Arousal	
Image Title	IAPS #	Mean	SD	Mean	SD
Woman	2039	3.81	1.15	3.44	1.91
Fingerprint	2206	3.91	1.51	3.56	2.18
Woman	2399	3.9	1.15	3.72	1.93
CryingFamily	2456	3.17	0.99	3.51	1.98
Man	2490	3.96	1.93	3.83	2.24
Smoking	2715	3.6	1.71	3.99	2.12
Cemetery	9000	2.81	1.65	3.9	2.12
Cemetery	9001	3.41	2.15	3.74	2.35
Family	9046	3.87	1.19	3.85	1.55
Puddle	9110	3.78	1.41	3.9	2.18
Garbage	9291	3.29	1.05	3.52	2.14
HomelessMan	9331	3.09	1.27	3.42	1.67
EmptyPool	9360	3.96	1.43	2.49	1.83
Dishes	9395	3.23	1.31	3.73	2.12
Cigarettes	9832	3.31	1.55	3.9	2.01

Male subjects	Low Valence - High Arousal				
		Valence		Arousal	
Image Title	IAPS #	Mean	SD	Mean	SD
NativeBoy	2730	3.4	2.92	6.6	2.25
Mutilation	3060	1.94	1.39	6.89	2.08
Mutilation	3068	2.47	1.92	6.44	2.46
Mutilation	3071	2.06	1.59	6.61	2.13
BurnVictim	3102	1.62	1.39	5.88	2.79
BabyTumor	3170	1.77	1.31	6.79	1.93
Surgery	3213	3.63	1.57	6.89	1.55
AimedGun	6260	2.53	1.63	7.1	1.9
DeadTiger	6415	2.81	1.63	5.86	2.27
Fire	8485	3.23	1.71	6.63	1.97
WarVictim	9250	2.85	1.47	6.5	1.66
Vomit	9322	2.64	1.36	5.8	2.08
Soldier	9410	1.96	1.56	6.38	2.26
Hanging	9413	2.23	1.32	6.06	2.35
Explosion	9940	1.91	1.29	7.37	2.03

Male subjects	**High Valence - Low Arousal**				
		Valence		**Arousal**	
Image Title	**IAPS #**	**Mean**	**SD**	**Mean**	**SD**
Butterfly	1604	6.4	1.31	3.17	1.98
Rabbit	1610	7.28	1.47	2.82	2.01
Fish	1900	6.4	1.67	3.04	2.07
Kid	2035	7.07	1.28	3.34	1.92
ChildCamera	2302	6.31	1.09	3.48	1.97
Girl	2304	6.42	1.23	3.17	1.85
Binoculars	2314	6.88	1.06	3.56	1.83
ThreeMen	2370	6.71	1.32	2.85	2.07
Couple	2501	6.33	1.86	2.67	2.3
Flowers	5200	6.96	1.62	3.46	2.06
Clouds	5551	6.79	1.49	3.28	2.01
Grain	5726	6.15	1.61	3.1	2.26
Nature	5760	7.69	1.28	2.77	2.16
Leaves	5800	6.21	1.83	2.54	2.22
Watermelon	7325	6.48	1.47	3.24	2.06

Male subjects	**High Valence - High Arousal**				
		Valence		**Arousal**	
Image Title	**IAPS #**	**Mean**	**SD**	**Mean**	**SD**
AttractiveFem	4007	7.7	1.53	7.39	1.3
Bikini	4090	7.64	1.26	7.18	1.3
AttractiveFem	4150	7.8	1.36	6.41	2.18
EroticFemale	4180	8.21	1.34	7.43	1.97
AttractiveFem	4250	8.39	0.93	7.02	2.02
EroticCouple	4607	7.99	1.09	7.19	1.88
EroticCouple	4659	7.7	1.64	7.43	1.8
Beach	5833	8.15	1.19	6.37	2.37
Sailing	8080	7.73	1.25	7.12	1.95
Hiker	8158	6.36	1.68	6.43	1.58
Bungee	8179	6.96	1.58	6.86	2.26
Skier	8190	8.13	1.29	6.41	2.6
Rafting	8370	7.67	1.19	6.46	2.22
Rollercoaster	8492	7.36	1.87	7.07	1.8
Money	8501	8.14	1.24	6.86	2

(Female Subjects)

Female subjects	Low Valence - Low Arousal				
		Valence		Arousal	
Image Title	IAPS #	Mean	SD	Mean	SD
Woman	2039	3.55	1.59	3.48	1.97
Boy	2280	3.97	1.73	3.96	2.02
Man	2490	2.74	1.51	4.06	1.77
ElderlyWoman	2590	3.46	2.24	3.86	1.93
Jail	2722	2.94	1.61	3.77	2.2
EroticFemale	4001	3.58	1.74	3.88	2.13
Prostitute	4233	3.89	1.7	3.43	1.85
EroticFemale	4235	3.67	1.82	3.97	2.44
Jail	6010	3.37	1.61	4.06	1.91
Cemetery	9001	2.82	1.88	3.6	2.27
Exhaust	9090	3.83	1.49	3.75	2.16
Puddle	9110	3.75	1.44	4.04	2.29
Woman	9190	3.63	1.58	4.08	1.71
Smoke	9280	2.69	1.47	4.05	2.35
Bridge	9472	3.9	1.31	4.04	2.01

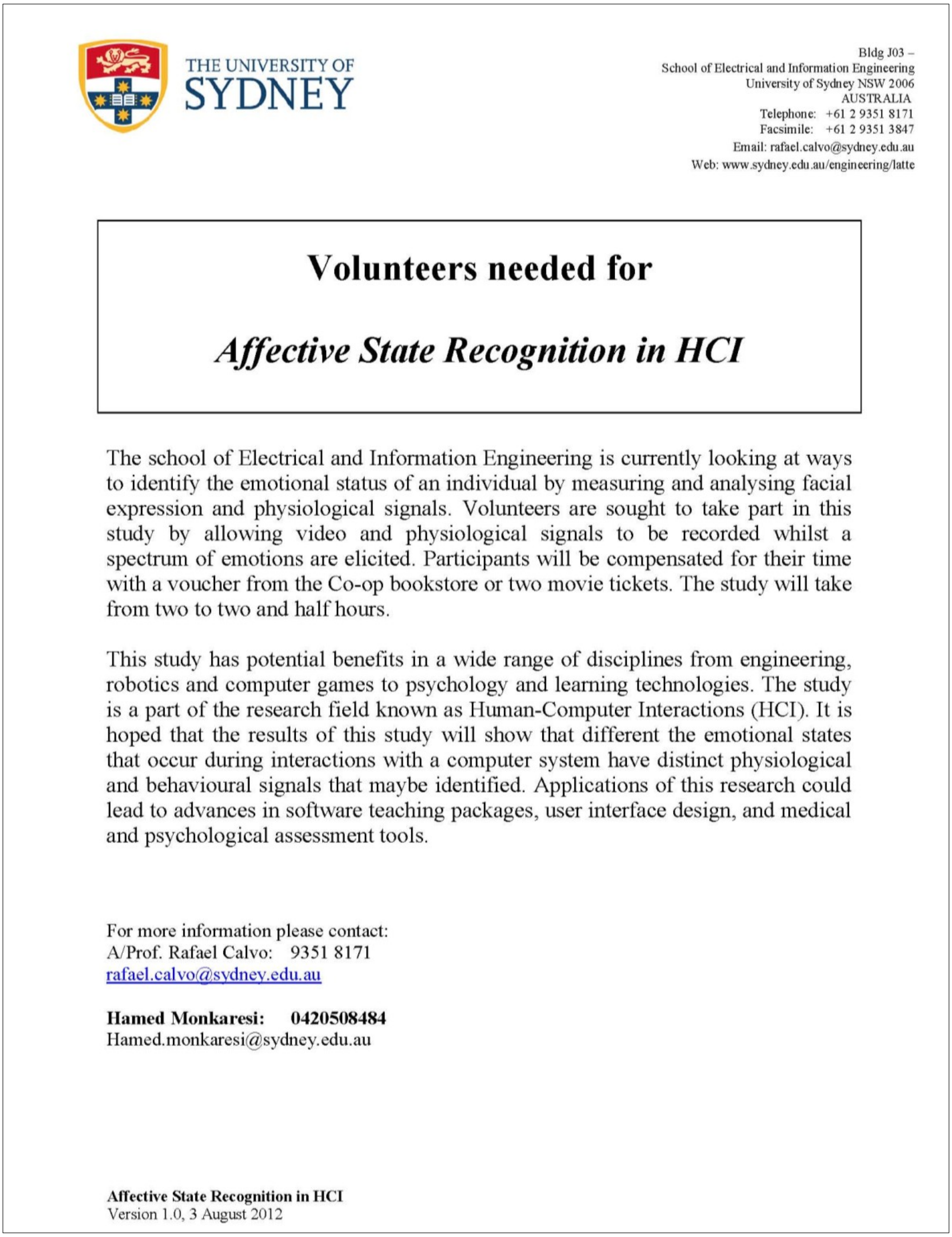
Female subjects	Low Valence - High Arousal				
		Valence		Arousal	
Image Title	IAPS #	Mean	SD	Mean	SD
Snake	1120	3.03	1.74	7.2	1.86
Spider	1201	2.93	1.81	6.87	2.09
Baby	2053	2.17	1.9	5.83	2.38
SadChildren	2703	1.59	0.87	5.81	2.47
Mutilation	3071	1.69	1.14	7.1	1.95
Mutilation	3080	1.33	0.75	7.61	1.81
DeadBody	3120	1.33	0.74	7.49	1.96
BabyTumor	3170	1.2	0.57	7.55	1.98
Soldier	6212	1.81	1.41	6.53	2.35
AimedGun	6230	2.06	1.59	7.56	1.96
Attack	6313	1.61	1.22	7.27	2.29
Attack	6520	1.59	1.01	7.12	1.72
Soldier	9410	1.2	0.58	7.54	1.78
DeadMan	9412	1.4	0.69	7.26	1.59
Hanging	9413	1.43	0.7	7.35	1.71

Female subjects	High Valence - Low Arousal				
		Valence		Arousal	
Image Title	**IAPS #**	**Mean**	**SD**	**Mean**	**SD**
Gannet	1450	6.87	1.54	2.8	1.87
Butterfly	1602	7.08	1.67	3.66	2.19
Rabbit	1610	8.39	0.91	3.33	2.36
Antelope	1620	7.95	1.19	3.49	2.36
Adult	2000	7.1	1.62	3.72	2.31
ChildCamera	2302	6.51	1.46	3.76	1.92
ThreeMen	2370	7.43	1.49	2.93	2.2
Kids	2388	8.1	1.15	3.73	2.46
Couple	2530	8.25	1.1	3.8	2.17
Flower	5000	7.59	1.63	2.9	1.92
Nature	5780	7.68	1.44	3.4	2.47
Flowers	5811	7.88	1.24	3.12	2.66
Clouds	5870	6.92	1.86	2.56	2.02
Ocean	7545	7.04	1.71	3.24	2.45
Violin	7900	6.5	1.69	2.37	2.01

Female subjects	High Valence - High Arousal				
		Valence		Arousal	
Image Title	**IAPS #**	**Mean**	**SD**	**Mean**	**SD**
Baby	2045	8.17	1.21	6.02	2.29
Bride	2209	7.95	1.46	5.91	2.4
Children	2216	7.85	1.18	6.29	2.13
EroticMale	4538	7.04	1.74	6.14	2.27
Wedding	4626	7.8	1.76	6.06	2.51
Romance	4640	7.64	1.85	5.94	2.46
EroticCouple	4643	6.73	1.68	5.92	2.19
EroticCouple	4698	6.38	1.56	6.58	1.79
SkyDivers	5621	7.8	1.54	7	2.13
Hiker	5629	7.15	1.51	6.52	2.04
Basketball	8001	7.46	1.28	6.62	1.83
Skier	8030	7.35	1.86	7.38	1.91
Rafting	8370	7.86	1.37	6.98	2.25
Rollercoaster	8492	7.11	2.49	7.48	1.51
Money	8501	7.67	1.97	6.02	2.5

APPENDIX B

Experiment Flayer



THE UNIVERSITY OF SYDNEY

Bldg J03 –
School of Electrical and Information Engineering
University of Sydney NSW 2006
AUSTRALIA
Telephone: +61 2 9351 8171
Facsimile: +61 2 9351 3847
Email: rafael.calvo@sydney.edu.au
Web: www.sydney.edu.au/engineering/latte

Volunteers needed for

Affective State Recognition in HCI

The school of Electrical and Information Engineering is currently looking at ways to identify the emotional status of an individual by measuring and analysing facial expression and physiological signals. Volunteers are sought to take part in this study by allowing video and physiological signals to be recorded whilst a spectrum of emotions are elicited. Participants will be compensated for their time with a voucher from the Co-op bookstore or two movie tickets. The study will take from two to two and half hours.

This study has potential benefits in a wide range of disciplines from engineering, robotics and computer games to psychology and learning technologies. The study is a part of the research field known as Human-Computer Interactions (HCI). It is hoped that the results of this study will show that different the emotional states that occur during interactions with a computer system have distinct physiological and behavioural signals that maybe identified. Applications of this research could lead to advances in software teaching packages, user interface design, and medical and psychological assessment tools.

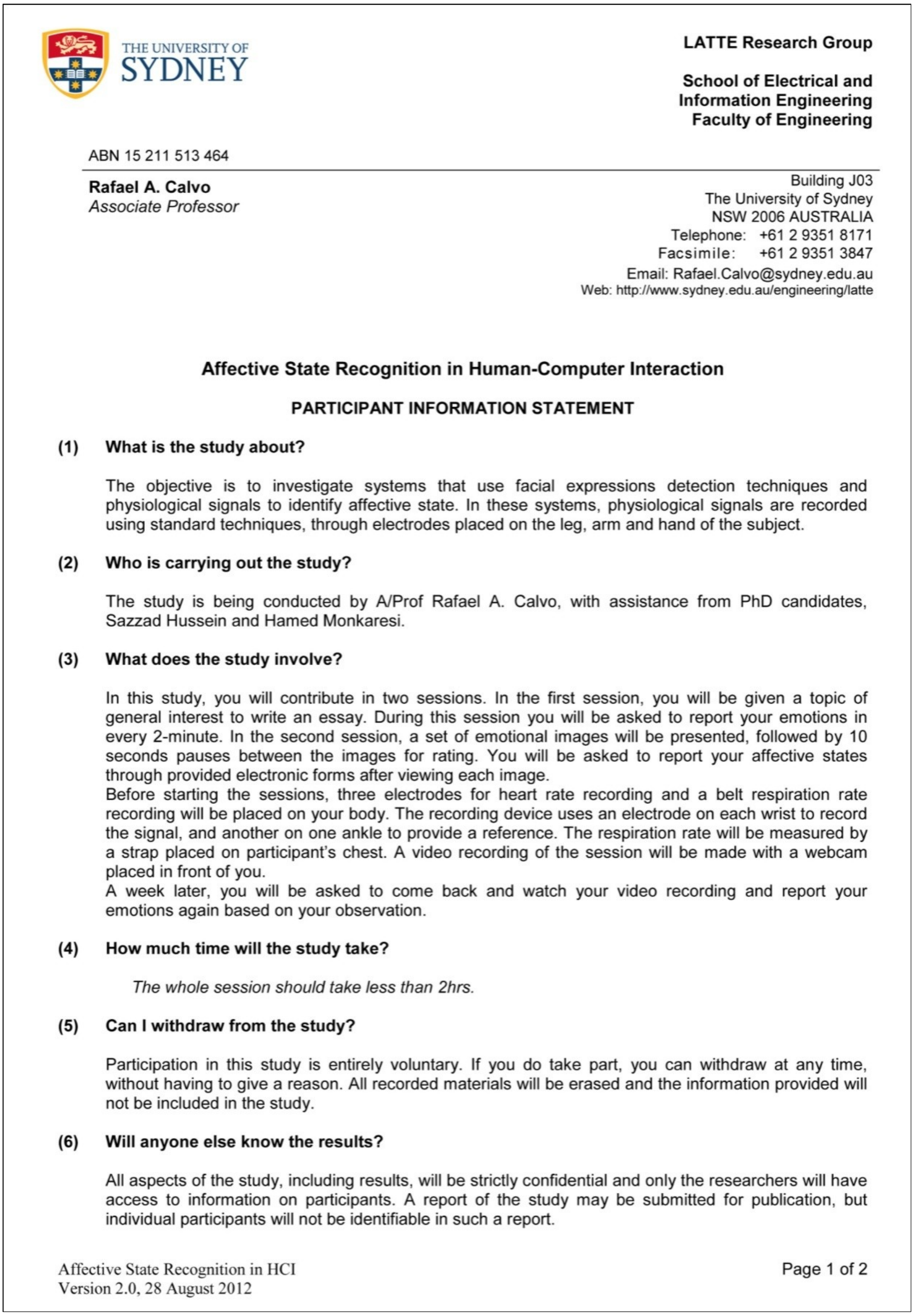
For more information please contact:
A/Prof. Rafael Calvo: 9351 8171
rafael.calvo@sydney.edu.au

Hamed Monkaresi: 0420508484
Hamed.monkaresi@sydney.edu.au

Affective State Recognition in HCI
Version 1.0, 3 August 2012

APPENDIX C

Consent Form



THE UNIVERSITY OF SYDNEY

LATTE Research Group

School of Electrical and
Information Engineering
Faculty of Engineering

ABN 15 211 513 464

Rafael A. Calvo
Associate Professor

Building J03
The University of Sydney
NSW 2006 AUSTRALIA
Telephone: +61 2 9351 8171
Facsimile: +61 2 9351 3847
Email: Rafael.Calvo@sydney.edu.au
Web: http://www.sydney.edu.au/engineering/latte

Affective State Recognition in Human-Computer Interaction

PARTICIPANT INFORMATION STATEMENT

(1) What is the study about?

The objective is to investigate systems that use facial expressions detection techniques and physiological signals to identify affective state. In these systems, physiological signals are recorded using standard techniques, through electrodes placed on the leg, arm and hand of the subject.

(2) Who is carrying out the study?

The study is being conducted by A/Prof Rafael A. Calvo, with assistance from PhD candidates, Sazzad Hussein and Hamed Monkaresi.

(3) What does the study involve?

In this study, you will contribute in two sessions. In the first session, you will be given a topic of general interest to write an essay. During this session you will be asked to report your emotions in every 2-minute. In the second session, a set of emotional images will be presented, followed by 10 seconds pauses between the images for rating. You will be asked to report your affective states through provided electronic forms after viewing each image.
Before starting the sessions, three electrodes for heart rate recording and a belt respiration rate recording will be placed on your body. The recording device uses an electrode on each wrist to record the signal, and another on one ankle to provide a reference. The respiration rate will be measured by a strap placed on participant's chest. A video recording of the session will be made with a webcam placed in front of you.
A week later, you will be asked to come back and watch your video recording and report your emotions again based on your observation.

(4) How much time will the study take?

The whole session should take less than 2hrs.

(5) Can I withdraw from the study?

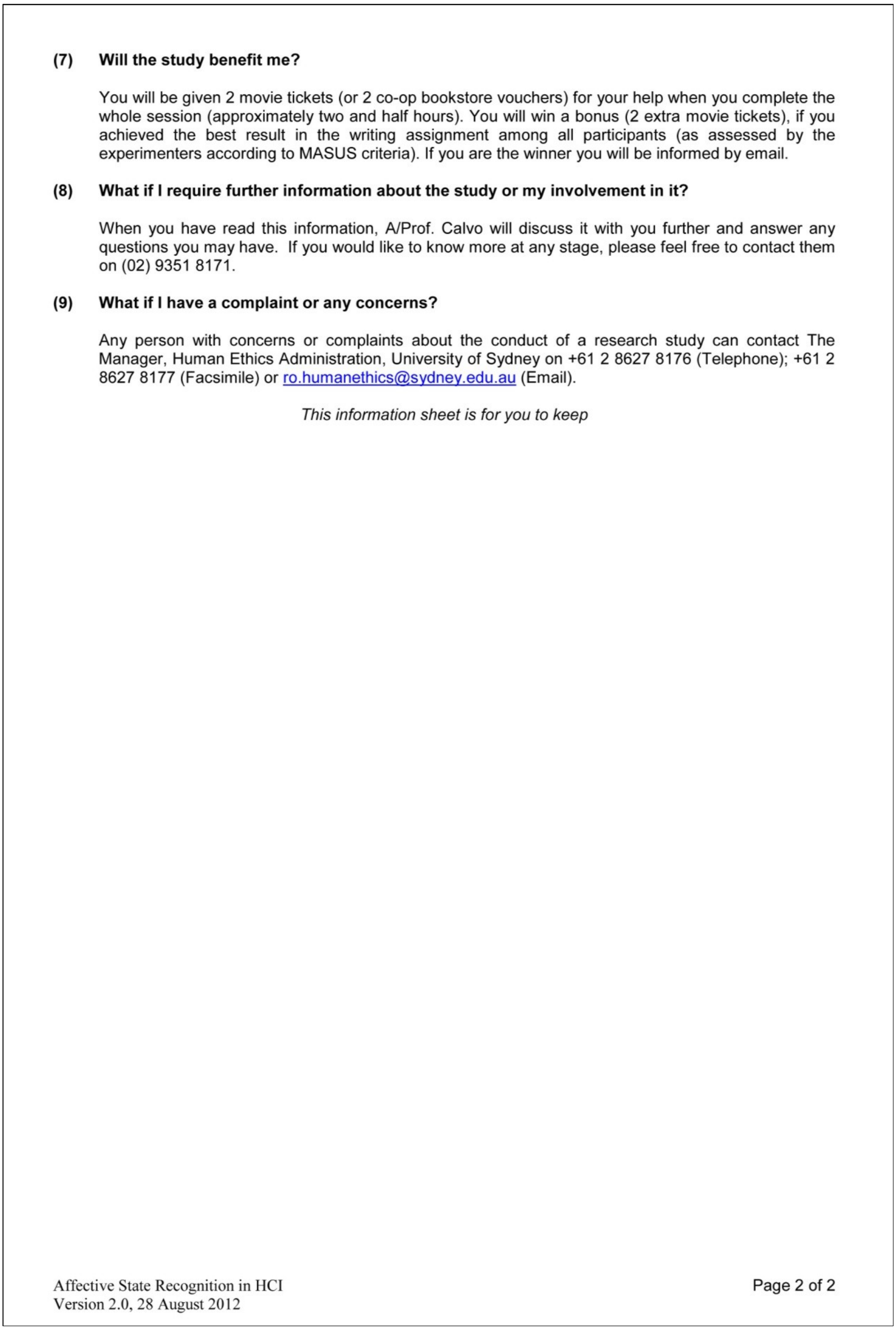
Participation in this study is entirely voluntary. If you do take part, you can withdraw at any time, without having to give a reason. All recorded materials will be erased and the information provided will not be included in the study.

(6) Will anyone else know the results?

All aspects of the study, including results, will be strictly confidential and only the researchers will have access to information on participants. A report of the study may be submitted for publication, but individual participants will not be identifiable in such a report.

Affective State Recognition in HCI
Version 2.0, 28 August 2012

Page 1 of 2



(7) Will the study benefit me?

You will be given 2 movie tickets (or 2 co-op bookstore vouchers) for your help when you complete the whole session (approximately two and half hours). You will win a bonus (2 extra movie tickets), if you achieved the best result in the writing assignment among all participants (as assessed by the experimenters according to MASUS criteria). If you are the winner you will be informed by email.

(8) What if I require further information about the study or my involvement in it?

When you have read this information, A/Prof. Calvo will discuss it with you further and answer any questions you may have. If you would like to know more at any stage, please feel free to contact them on (02) 9351 8171.

(9) What if I have a complaint or any concerns?

Any person with concerns or complaints about the conduct of a research study can contact The Manager, Human Ethics Administration, University of Sydney on +61 2 8627 8176 (Telephone); +61 2 8627 8177 (Facsimile) or ro.humanethics@sydney.edu.au (Email).

This information sheet is for you to keep