Intelligent Emotion Recognition from Facial and Whole-body Expressions using Adaptive Ensemble Models

Yang Zhang

A thesis submitted to the

University of Northumbria at Newcastle

for the degree of

Doctor of Philosophy

Department of Computer Science and Digital Technologies,
Faculty of Engineering and Environment,
Mar 2015

Acknowledgement

I would like to express my gratitude to my supervisor, Dr Li Zhang, whose guidance and support were paramount to this work. She patiently provided the constructive advice and unflagging encouragement for me. I would also like to thank my dear wife, Chunmin Han. Without her love and spiritual support, I would never go this far. I would also like to thank my parents for their constant support and encouragement throughout all my student life. Also, I really appreciate all the discussions conducted with Prof. Alamgir Hossain and other fellow PhD researchers in the Faculty. Last but not least, I would also like to thank the internal and external examiners for their time spent on reading the thesis and providing helpful suggestions.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the Faculty Ethics Committee on 11/2011.

I declare that the Word Count of this Thesis is 36000 words.

Name: Yang Zhang

Signature:

Abstract

Automatic emotion recognition has been widely studied and applied to various computer vision tasks (e.g. health monitoring, driver state surveillance, personalized learning, and security monitoring). With the great potential provided by current advanced 3D scanners technology (e.g. the Kinect), we shed light on robust emotion recognition based one users' facial and whole-body expressions. As revealed by recent psychological and behavioral research, facial expressions are good in communicating categorical emotions (e.g. happy, sad, surprise, etc.), while bodily expressions could contribute more to the perception of dimensional emotional states (e.g. the arousal and valence dimensions). Thus, we propose two novel emotion recognition systems respectively applying adaptive ensemble classification and regression models respectively based on the facial and bodily modalities.

The proposed real-time 3D facial Action Unit (AU) intensity estimation and emotion recognition system automatically selects 16 motion-based facial feature sets to estimate the intensities of 16 diagnostic AUs. Then a set of six novel adaptive ensemble classifiers are proposed for robust classification of the six basic emotions and the detection of newly arrived unseen novel emotion classes (emotions that are not included in the training set). In both offline-line and on-line real-time evaluation, the system shows the highest recognition accuracy in comparison with other related work and flexibility and good adaptation for newly arrived novel emotion detection(e.g. 'contempt' which is not included in the six basic emotions). The second system focuses on continuous and dimensional affect prediction from users' bodily expressions using adaptive regression. Both static posture and dynamic motion bodily features are extracted and subsequently selected by a Genetic Algorithm to identify their most discriminative combinations for both valence and arousal dimensions. Then an adaptive ensemble regression model is proposed to robustly map subjects' emotional states onto a continuous arousal-valence affective space using the identified feature subsets. Experimental results show that the proposed system outperforms other benchmark models and achieves promising performance compared to other state-of-the-art research reported in the literature. Furthermore, we also propose a novel semi-feature level bimodal fusion framework that integrates both facial and bodily information together to draw a more comprehensive and robust dimensional interpretation of subjects' emotional states. By combining the optimal discriminative bodily features and the derived AU intensities as inputs, the proposed adaptive ensemble regression model achieves remarkable improvements in comparison to solely applying the bodily features.

Contents

List of Figures

List of Tables

List of Abbreviations

AANN	Auto-Associative Neural Network
AU	Action Unit
BLSTM-NN	Bidirectional Long Short-Term Memory Neural Network
BPNN	feedforward Neural Network with Backpropagation
CMTNN	Complementary Neural Network
CORR	Pearson correlation coefficient
EDA	Estimation Distribution Algorithm
FACS	Facial Action Coding System
FAPUs	facial animation parameter units
FLD	Fisher Linear Discriminant
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
HCRFs	hidden conditional random fields
HMM	hidden Markov model
KLT	Karhunen-LoeveTransform
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory Neural Network
MDA	mixture discriminant analysis
MI	mutual information
MLR	multivariate Logistic Regression
MRBF	Median Radial Basis Function
mRMR	minimal-redundancy-maximal-relevance
MSE	Mean Squared Error
NCBF	normalized cut-based filter
NN	Neural Network

PCA	Principle Component Analysis
RBF	Radial Basis Function
SFFS	Sequential Forward Floating Selection
SR	Sparse Representation classifier
SVM	Support Vector Machine
SVR	Support Vector Regression

List of Publications

Conference Papers:

- Zhang, Y., Zhang, L., Hossain, A., Jiang, Y. and Spencer, L. (2012). Towards the Development of a Multimodal Framework for Intelligent Affect Detection. *Proceedings of the 6th Conference on Software, Knowledge, Information Management and Applications*, SKIMA 2012.

- Zhang, Y., Zhang, L. and Hossain, A. (2013). Multimodal Intelligent Affect Detection with Kinect. *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems,* AAMAS 2013.

- Wang, Y., Angelova, M. and Zhang, Y. (2013). A framework for density weighted kernel fuzzy c-means on gene expression data. *Advances in Intelligent Systems and Computing*, Volume 212, pp 453-461.

- Zhang, Y. and Zhang, L. (2015), Semi-feature Level Fusion for Bimodal Affect Regression Based on Facial and Bodily Expressions. Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems.

Journal Papers:

- Zhang, Y., Zhang, L. and Hossain, A. (2015). Adaptive 3D Facial Action Intensity Estimation and Emotion Recognition. *Expert Systems with Applications*, 42 (3), 1446–1464.

- (Submitted) Zhang, Y. and Zhang, L. (2015). Intelligent Affect Regression for Whole-Body Expressions Using Adaptive Ensemble Models. *Expert Systems with Applications*, ELSEVIER.

Chapter 1 Introduction

1.1 Background

In recent years, ubiquitous computer and information technology has become a more and more indispensable part of our everyday life. It also drives innovations in agent-based interface development. Also, as an important aspect of human life, emotion and affect help us to express and perceive our goals, feelings and intentions, subtly impacting our daily activities such as learning, decision making and interpersonal communication. Thus, in the era where computer technology and human life have become extremely interwoven, automatic emotion recognition has become a new hotspot of AI research since the role played by affect in human life and everyday functioning is well recognized and studied (Izard et al., 2000).

Endowing machines with emotion intelligence not only greatly benefits natural Human-Computer Interaction, but also shows great potential to be applied in a wide variety of applications, such as personalized learning (D'Mello & Graesser, 2010), health monitoring (Lucey et al., 2009), customer services (Zeng et al., 2009), anomalous event detection (Ryan et al., 2009), intelligent robotics (Fellous & Arbib, 2005), and interactive computer entertainment (Savva et al., 2012; G'Mussel & Hewig, 2013). Emotional information can be expressed and perceived through a wide range of non-verbal channels, such as face, voice, text, and bodily expressions. In this research we particularly shed light on the facial and bodily modalities, because of the importance and potential of those modalities to human affective behavior interpretation revealed by recent research (e.g. Kleinsmith & Bianchi-Berthouze, 2013; Chavan & Kulkarni, 2013).

1.2 The role of facial expressions and challenging problems

Facial expressions are the facial changes caused by underlining muscle movements, and in response to a person’s internal feelings, intentions, emotional states, or social communications (Tian, 2005). Facial expression analysis was primarily a research subject for behavioral scientists and psychologists, since the seminal research by Darwin (1872). A milestone in facial expression research is the postulation of six basic emotions (i.e. happiness, surprise, fear, anger, sadness, and disgust), each of which possesses a distinctive content together with a unique prototype facial expression and is claimed to be universal across cultures and human ethnicities (Ekman & Friesen, 1971).

Suwa et al. (1978) started a new era for automatic facial expression analysis. In their preliminary investigation, they attempted to automatically analyze facial expressions from an image sequence by tracking the motion of twenty identified facial landmarks. After that, the field of automatic facial expression recognition has drawn ever-increasing attention. The last decade has witnessed significant progress in the related areas (e.g. Pantic & Patras, 2006; Cohn et al., 2009; Sorci & Thiran, 2010; Kappas, 2010; Tsalakanidou & Malassiotis, 2010; Zhang, 2011; Valstar & Pantic, 2012; Koelstra et al., 2010; Savran et al., 2012; Wang et al., 2006; Mpiperis, 2008; Zhang et al., 2013; Owusu et al., 2014; Rao et al., 2011). Thus, the importance and role of facial expressions in the expression and perception of emotions have been widely studied and accepted in both cognitive neuroscience and computer science.

Most existing facial emotion recognition systems, however, either only considered static facial features, or were limited to 2D models. They have not fully considered dynamic information of facial movements that are relatively subject-independent and may play a critical role in interpreting emotions, thus are not robust enough for challenging real-life recognition tasks with subject variation, head movement and illumination changes. Moreover, a good facial emotion recognition system is also expected to be well capable of detecting the arrival of novel emotion classes (e.g. compound emotions or other new emotions that do not belong to the six basic emotion categories mentioned in the training set). However, there is lack of systematic research for the effective detection of novel emotions.

To address these challenges, we present a real-time 3D facial Action Unit (AU) intensity estimation and emotion recognition system. We first of all extract dynamic motion-based facial features to robustly estimate the intensities of 16 selected Action Units (AUs) using Neural Networks (NNs) and Support Vector Regressors (SVRs). Subsequently, a set of six novel adaptive ensemble classifiers is proposed for the detection of six basic expressions and any newly arrived novel emotion classes. The details are presented in Chapter 3.

1.3 The role of bodily expressions and challenging problems

In 1872, Darwin presented the first rigorous evidence for the expression of emotions through the body (Darwin, 1872). A series of bodily behaviors specific to certain emotional categories was found in his work, many of which are now regarded as basic emotions (e.g. anger, disgust and surprise). In the following century, the role of body language in the expression and perception of emotions has also been well revealed by many other researchers (e.g. Wallbott, 1998; Montepare et al., 1999; Van den Stock et al., 2007; de Gelder, 2009). Compared to the booming research on automatic facial expression recognition in the last decade, only recently there have been fewer automatic systems that are able to detect emotions based on the bodily modality (e.g. Bernhardt & Robinson, 2007; Kleinsmith et al., 2011; Kleinsmith & Bianchi-Berthouze, 2013). This may be attributed to the complexity of the body itself and the lack of well acknowledged coding models for the body as there are for the face (e.g. the well-established Facial Action Coding System (Ekman et al., 2002)).

Recent studies in cognitive neuroscience (de Gelder et al., 2003; Van den Stock et al., 2007) have emphasized that body posture could be the influencing factor over facial expression in cases of incongruent affective displays, and for the discrimination between some emotional states in particular, such as fear and anger, more attention needs to be paid to the bodily display. Furthermore, a longstanding controversy in cognitive science has concerned whether emotions are better conceptualized in the form of discrete categories (e.g. happy and sad), or continuous dimensions (e.g. valence and arousal) (Hamann, 2012). According to Ekman & Friesen (1967), compared to the face, which is considered to be the foremost modality for expressing discrete emotion categories, the body may perform better for communicating affective dimensions. Recent research (Kleinsmith & Bianchi-Berthouze, 2013) also indicates that by the combination of discrete emotion labels and continuous dimension levels, a more complete and systematic description of the emotional state could be obtained. These highlight the importance of developing a dimensional emotion recognition system based on bodily expressions.

More importantly, current neuroscience studies (Vania et al., 1990; Giese & Poggio, 2003; Lange & Lappe, 2007) indicate that our brain utilizes two separate pathways for the recognition of biological information from bodily expressions, one for form information (e.g. a specific configuration of a posture), and the other for motion information (e.g. velocity, acceleration, and frequency). According to Atkinson et al. (2007), both form and motion bodily signals make their own contributions to affect perception of human behavior. A number of recent developments in computer science (e.g. Roether et al., 2009; Kleinsmith et al., 2011) further prove that both of them are useful and important for automatic emotion prediction from bodily expressions. Body form and motion information complement each other in conveying emotions, however, they may also become partially redundant or inconsistent in some cases (Kleinsmith & Bianchi-Berthouze, 2013). Thus, it is also significant to identify the roles of both body form and movement information in the automatic regression of different affective dimensions.

Thus, we also aim to address the problem of continuous regression of subjects' emotional states in a valence and arousal space based on their whole-body expressions. I.e. the proposed system is able to robustly map subjects' emotional states to a two-dimensional coordinate space spanned by arousal and valence, where each value ranges between -1 and 1. We systematically extract users' static and dynamic bodily features and conduct feature selection using the Genetic Algorithm (GA) based optimization. An ensemble regression model with great adaptability is also proposed to deal with continuous prediction of subjects' affective dimensions. The details are presented in Chapter 4.

1.4 Research contribution

Within the research area of affective computing and machine learning, the contributions of this thesis are threefold.

1. First of all, we propose an automatic system for real-life 3D AU intensity estimation and categorical facial expression recognition with novel emotion detection.

    - We extract dynamic motion-based facial features (e.g. the elongation of mouth) rather than static features (e.g. the width of mouth) to estimate AU intensities because of the following. Static features could change a lot between different subjects, whereas the motion-based features are caused by underlying facial muscle movements which bear anatomically similar muscle tension behavior among different subjects for the expression of the six basic emotions (Ekman et al., 2002), and thus are relatively universal and subject-independent, and contain comparatively richer emotional information. Therefore they are employed in this research for facial expression representations.
    - A minimal-redundancy-maximal-relevance criterion (mRMR) based automatic feature selection is proposed to identify the most discriminative and informative feature sets for AU intensity estimation. Compared with the manual feature selection conducted based on facial muscle anatomical and

FACS knowledge, the mRMR-based optimization yields comparable performance for the intensity estimation of the 16 selected AUs.

- We also propose a set of six novel adaptive ensemble classifiers to robustly differentiate between the six basic emotions and identify newly arrived unseen novel emotion categories. Each ensemble model employs a special type of Neural Network, i.e. Complementary Neural Network, as the base classifier, which is able to provide uncertainty measure of its classification performance. We consider the following idea for novel class detection. Instances within the same emotion categories should be close to each other whereas those from different categories should indicate great distinction to each other. Therefore, a distance-based clustering and the uncertainty measures of the base Complementary Neural Network classifiers are used to inform the arrival of novel unseen emotion classes. The proposed ensemble models achieve 92.2% average accuracy and consistently outperform other single Support Vector Machine classifiers employed in this research and other related research reported in the literature when evaluated with the Bosphorus database (Savran et al., 2008).
- The proposed system is also evaluated with real-time emotion detection tasks contributed by real human subjects. The system achieves comparable accuracy (84%) in comparison to the results gained from the evaluation using database images. It also shows great adaptation and robustness for newly arrived novel emotion class detection with ≥70% accuracy. The system is therefore proved to be effective in dealing with challenging real-life emotion recognition tasks.

2. Equally importantly, the second system proposed in this thesis aims to address the problem of continuous and dimensional interpretation of users' emotional states based on their whole-body expressions. I.e. subjects' emotional states are mapped to a two-dimensional coordinate space spanned by arousal and valence, where each

value ranges between -1 and 1.

- We systematically consider and extract users’ static and dynamic bodily features. The GA algorithm is then employed to conduct feature selection and identify their most optimal discriminative combinations for affective dimensional regression. We also examine how both static and dynamic features perform for the regression of each affective dimension.
- An ensemble regression model with great adaptability is also proposed to robustly predict users’ continuous affective dimensions in the valence and arousal space using whole-body expressions. The proposed ensemble model with Support Vector Regressors as the base regressors achieves the best performance and outperforms single model based methods and other related research reported in the literature. Furthermore, it also employs a stand-by regressor to better deal with newly arrived unseen bodily expressions and data stream regression.
- Continuous and dimensional affective annotation is inherently a challenging task. We present a novel annotation method based on inter-annotator correlations and mean value differences to effectively fuse multiple annotations to build ground truth for system evaluation.

3. Furthermore, based on the empirical findings of the above two systems, we proposed a semi-feature level fusion framework that effectively combines affective information from both the facial and bodily modalities to boost the performance of the dimensional affect recognition.

   - The semi-feature level fusion is realized by concatenating the derived AU intensities and the optimal discriminative bodily features into a merged feature vector which is subsequently employed as inputs to ensemble regressors, and shows significant performance improvements in comparison to sole applying the bodily features.

1.5 Thesis outline

Figure 1-1 shows the overall system architecture. The rest of this thesis is organized as follows.

Chapter 2 provides a thorough review of the literature from different disciplines. It starts with a brief discussion of diverse emotion theories followed by an introduction to the Facial Action Coding System. Then we survey related work in the field of automatic affect recognition from both facial and bodily expressions, and identify representative research challenges.

Chapter 3 presents the methodology and implementation of the proposed facial expression recognition system, including facial geometric feature tracking, mRMR-based feature selection, AU intensity estimation and facial expression recognition with novel emotion detection using adaptive ensemble classifiers. Subsequently, we conduct extensive experiments with both on-line and off-line evaluations for AU intensity estimation and emotion recognition.

Chapter 4 presents the proposed continuous and dimensional affect recognition system based on whole-body expressions. We first discuss feature extraction from whole-body expressions and automatic feature selection using the GA optimization. Then, the proposed adaptive ensemble regression model for continuous and dimensional affective regression is discussed in detail. We subsequently present the process of data collection and affective annotation method for system evaluation, as well as experiments and discussions.

Chapter 5 explores the modality fusion for dimensional affect recognition. We first of all review state-of-the-art developments on multimodal emotion recognition. The proposed semi-feature level framework is presented subsequently, together with experimental results and evaluation.

Chapter 6 summarizes the contributions and identifies future work.

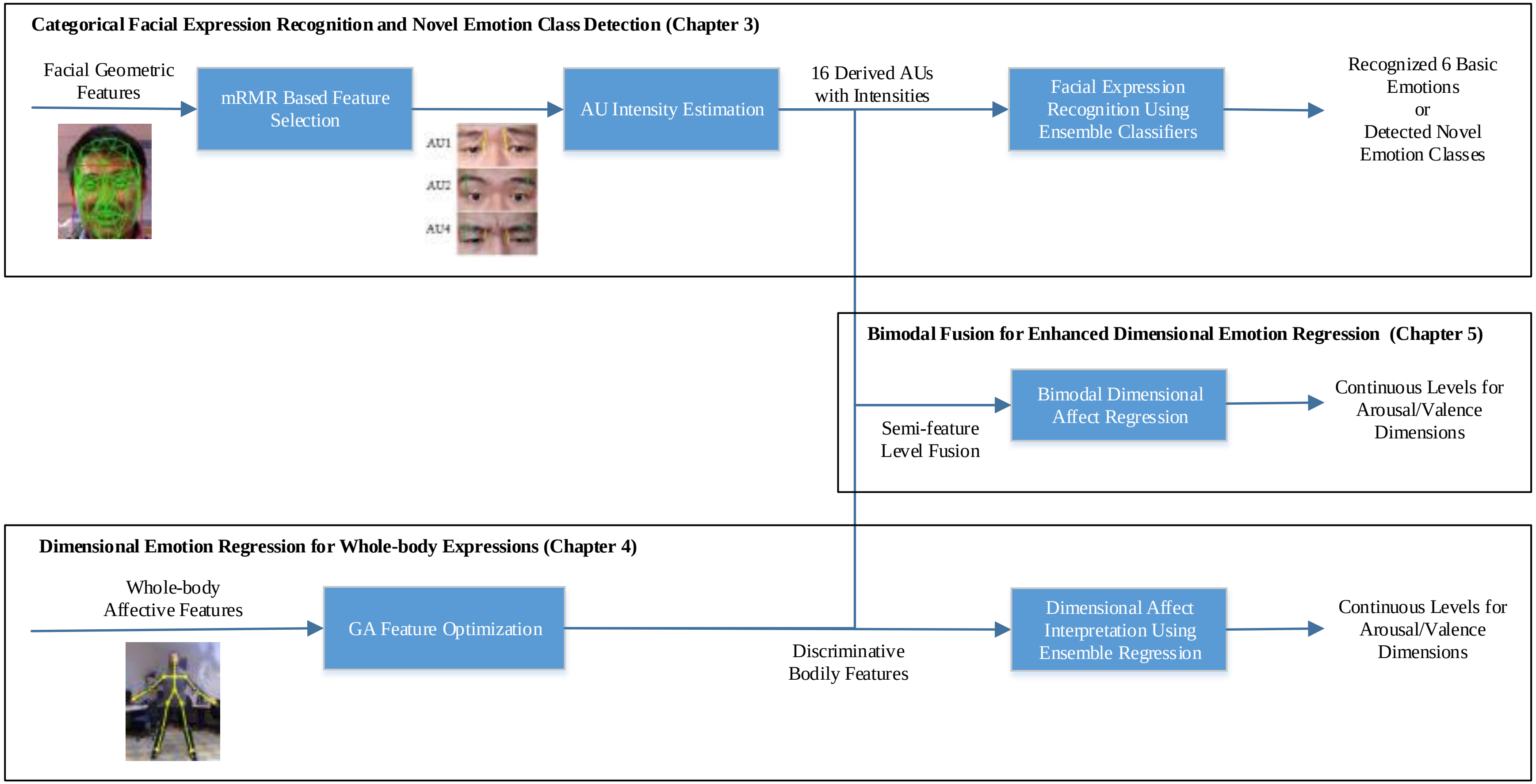

Categorical Facial Expression Recognition and Novel Emotion Class Detection (Chapter 3)
Facial Geometric Features
mRMR Based Feature Selection
AU1
AU2
AU4
AU Intensity Estimation
16 Derived AUs with Intensities
Facial Expression Recognition Using Ensemble Classifiers
Recognized 6 Basic Emotions or Detected Novel Emotion Classes
Bimodal Fusion for Enhanced Dimensional Emotion Regression (Chapter 5)
Bimodal Dimensional Affect Regression
Semi-feature Level Fusion
Continuous Levels for Arousal/Valence Dimensions
Dimensional Emotion Regression for Whole-body Expressions (Chapter 4)
Whole-body Affective Features
GA Feature Optimization
Discriminative Bodily Features
Dimensional Affect Interpretation Using Ensemble Regression
Continuous Levels for Arousal/Valence Dimensions

Figure 1-1 The overall system architecture

Chapter 2 Related work

In this chapter, we firstly provide a succinct overview on the conceptualization of emotions universally acknowledged by psychologists and behavioral scientists. Then we introduce some essential psychological theories and domain knowledge for the face and body in affect communication respectively. Afterwards, we discuss existing research work in the field of affective computing and conduct a concise survey on state-of-the-art emotion recognition developments.

2.1 Modelling of emotions - discrete vs continuous

In the literature of psychology, there are mainly two different approaches to structure and differentiate between different emotional states: **discrete categories** and **continuous dimensions**. The discrete model argues that the affective state is able to be represented by a number of prototypical emotions or their mixtures. This model has been well adopted and promoted by Ekman et al. (2002) and Izard (1994). According to their studies, there exists a series of basic emotions that can be expressed through corresponding prototypical facial expressions. For example, Figure 2-1 (a) shows facial expressions for the six basic emotions (i.e. happiness, surprise, fear, anger, sadness, and disgust) (Ekman & Friesen, 1967).

The continuous model argues that emotions are able to be described by certain continuous attributes, and the affective state of each participant could be placed within a continuous low-dimensional space. A representative model proposed by Posner et al. (2005) employed two orthogonal dimensions: valence and arousal. The valence dimension describes the level of pleasure of an emotion, and it ranges from negative unpleasant feelings to positive pleasant feelings. The arousal dimension refers to the intensity of the emotional experience, and it ranges from apathetic sleepiness to frantic excitement. Figure 2-1 (b) illustrates the two dimensional emotion model and the distributions of some identified emotion categories (Posner et al., 2005; Breazeal, 2003). The dimensional model could be a more flexible and effective way to interpret emotions, especially in the cases of (1) no clear categorical description available for an emotional state; (2) bodily expression-based continuous emotion communication, as indicated by Ekman & Friesen (1967). Therefore, we also borrow the dimensional model of valence and arousal for the automatic interpretation of emotional bodily expressions and maps emotional bodily behaviors to this two-dimensional continuous space in this work.

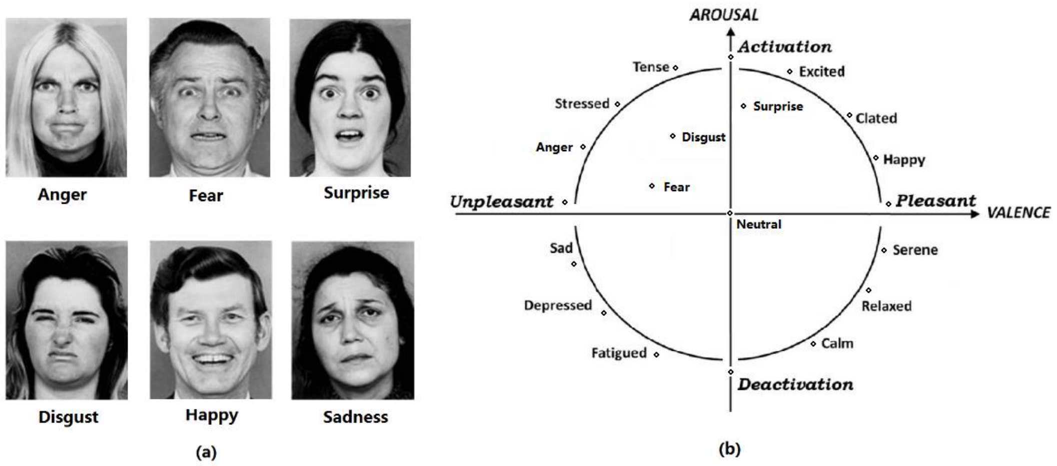


Figure 2-1 a. Facial expressions for the six basic emotions (Ekman & Friesen, 1967)

b. The arousal-valence two dimensional model and the distributions of some standard emotions (Posner et al. (2005) and Breazeal (2003))

2.2 FACS and related facial muscle anatomy

Facial Action Coding System (FACS) (Ekman et al., 2002) is widely used for facial emotion research in both psychology and computer science fields. It is an objective and comprehensive system based on the research of experimental psychologists, which aims to provide human expert observers with objective measures of facial activities. In the field of psychology and behavioral science, FACS represents the most recognized standard for facial expression analysis and measurement. A total of 46 facial Action Units (AUs) is defined to represent all possible subtle changes in muscle activations caused by emotional expressions, conversational and other facial behaviors. The original coding rules are generated based on visually discernible facial appearance changes observed from a large amount of images. According to FACS, every facial expression can be decomposed and represented by one AU or a combination of AUs. The intensity of an AU can be scored on a five-point ordinal level, from A to E (see Figure 2-2). The definitions of these levels are provided in the following. Level A refers to a trace of an action. Level B indicates slight evidence. Level C describes pronounced or marked evidence. Level D represents severe or extreme actions with Level E indicating maximum evidence. Each intensity level refers to a range of appearance changes.

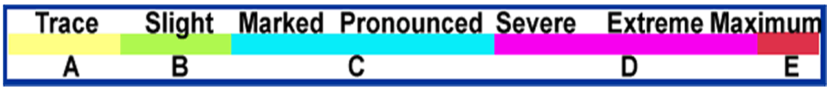
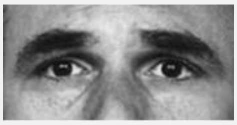
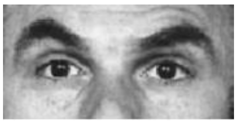
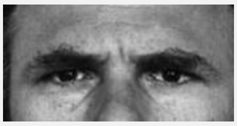
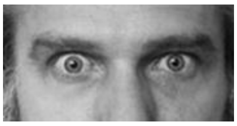
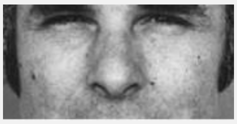
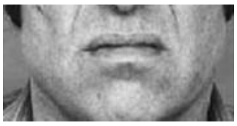
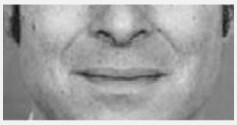
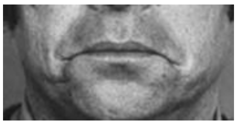
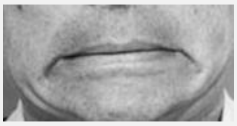
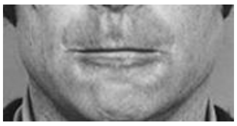
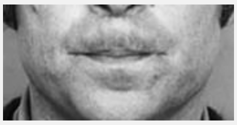
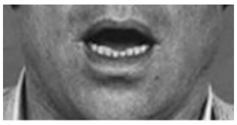


Figure 2-2 Five levels for AU intensity scores (Ekman et al., 2002)

In FACS, each AU is anatomically related to the contraction and relaxation of one or a specific set out of the 17 facial muscles. Each muscle is innervated by a specific facial nerve and contributes to one or a number of AU(s), while a single AU can also be associated with more than one muscles. These muscles are related to each other dynamically and spatially, generating every subtle change of Action Units and enabling coherent and consistent facial expressions (Ekman et al., 2002). Table 2-1 summarizes some AU examples, their associated facial muscles and corresponding emotions. The possible interpretations of emotions pertaining to each AU are also provided. By noticing specific changes of corresponding AUs, one can visually perceive and recognize each subtle facial expression.

Table 2-1 AUs, associated facial muscles, and corresponding expressions (Ekman et al., 2002)

AU Number and Name	***Facial Muscles***	***Possible Expressions***	***Example pictures***
AU1 Inner Brow Raiser	Frontalis, Pars Medialis	Sadness	
AU2 outer Brow Raiser	Frontalis, Pars Lateralis	Anger, Surprise	
AU4 Brow lower	Procerus	Anger, Anxiety, Pain	
AU5 Upper Lid Raiser	Levator Palpebrae Superioris	Fear, Surprise, Anger	
AU6 Cheek Raiser	Orbicularis, Oculi, Pars, Orbitalis	Happy	
AU10 Upper lip Raiser	Levator Labii Superioris	Disgust	
AU12 Lip Corner Puller	Zygomaticus Major	Happy	
AU15 Lip Corner Depressor	Triangularis	Sadness, Unsatisfying	
AU20 lip Stretcher	Risorius	Fear	
AU23 Lip Tightner	Orbicularis Oris	Anger	
AU24 Lip Pressor	Orbicularis Oris	Anxiety	
AU26 jaw Drop	Masetter	Surprise	

2.3 Body form vs motion information in conveying emotions

Bodily expressions are composed of two aspects: form and movement information. The former, better known as body posture, is usually defined as the static configuration of body parts, such as head pose, hand gesture, as well as trunk, arm and leg configuration. The latter normally refers to dynamic body motion, which can be quantified in terms of velocity, acceleration, amplitude, frequency, etc.

The importance of body posture in conveying emotion was first investigated by Darwin (1872), and has been widely explored in the following century (e.g. Bull, 1987; Wallbott, 1998). Figure 2-3 lists some archetypal example body gestures depicting some basic emotions in the work of Darwin (1872), i.e. disgust, anger, helplessness, and surprise. More recent research (Coulson, 2004) conducted a more systematic analysis on the effect of a variety of body posture features. Especially, in this research, we also observe that 'anger' is able to be expressed by postures such as "arms raised forward and upward, head bent back, and no backward chest bend" (e.g. an angry protestor), and 'happiness' is usually represented by postures such as "arms raised above shoulder, straight at elbow, head bent back, and no forward chest movement" (e.g. an excited football fan).

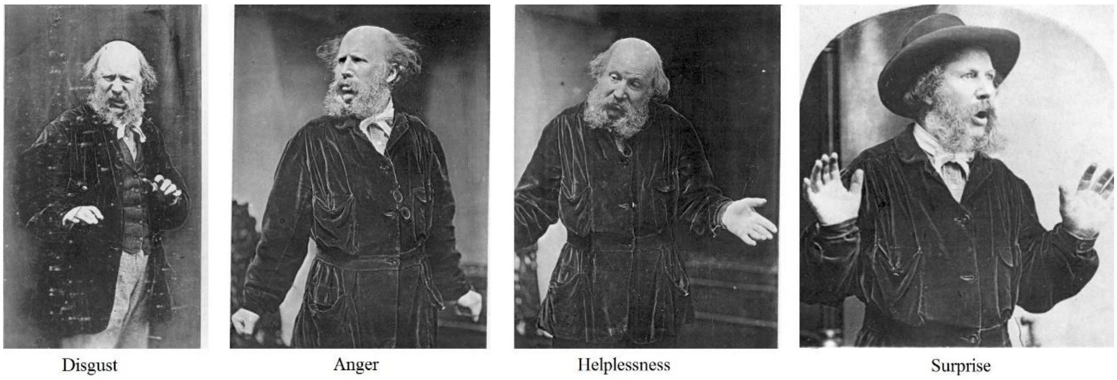


Figure 2-3 Archetypal gestures associated with some basic emotions (Darwin, 1872)

Similarly, the significance of dynamic body motion in emotional communication has also been revealed in a variety of contexts, such as children's music (Boone & Cunningham, 2001), dance (Camurri et al., 2003), interactive dialogues

(Clarke et al., 2005), and everyday activities such as knocking or drinking behaviors (Pollick et al., 2001). A series of movement qualities such as speed, jerkiness, and rhythm has been confirmed to be effective indicators of emotions. These findings are consistent with previous predictions by Scherer and Wallbott (1990), who indicated that it is possible to detect emotions by bodily expressions through changes in the speed, rhythm and fluidity of movements. This evidence provides forceful sufficient support for automatic emotion perception from body movements.

More recently, the work by Atkinson et al. (2004) and their more comprehensive follow-up study (Atkinson et al., 2007) concluded that both static form and dynamic motion signals were utilized to perceive emotions from bodily expressions. They further pointed out that body form and movement information can provide distinctive contributions to the perception of different emotion categories (e.g. body form information usually plays a greater role in the perception of 'fear' and 'disgust' than motion information). Roether et al. (2009) also indicated that analyzing posture cues can help to discriminate between emotions that are associated with similar dynamic features (e.g. 'happiness' and 'anger' could have similar features in terms of limb motion, but they can be distinguished through the analysis of posture cues). These studies not only revealed the importance of both body form and motion in conveying emotions, but also highlighted the significance of identifying their respective roles in recognizing specific emotion categories or affective dimensions.

2.4 Automatic emotion recognition from facial expressions

There has been extensive research focusing on automatic facial emotion recognition. Overall, the existing approaches in the area can be categorized into two groups: **static** and **dynamic** features based.

The static feature based systems normally focused on recognizing emotional facial expressions by observing representative facial geometric (e.g. points or shapes of facial components) or appearance features (e.g. facial wrinkles, furrows or bulges) statically and directly from the image data. For example, Soyel & Demirel (2007) extracted six characteristic distance features from the distribution of 11 manually labelled facial feature points in a 3D facial model, and then employed them as inputs to a Neural Network classifier for the recognition of the six basic emotions. Rao et al. (2011) extracted grey pixel features from eye and mouth regions, and then used Auto-Associative Neural Network (AANN) models to capture the distribution of the extracted features. Their system achieved an 87% average accuracy for the recognition of anger, fear, happiness, and sadness from video inputs. Another representative result was obtained by Tang & Huang (2008). They extracted 96 static distance and slope features from a cropped 3D face mesh model with 87 landmark points. The derived features were also normalized by facial animation parameter units (FAPUs) to ensure their person-independence. By using multi-class Support Vector Machine (SVM) classifiers, an average accuracy of 87.1% was achieved for the recognition of the six basic emotions, with the highest classification rate of 99.2% obtained for surprise. Mahoor et al. (2011) employed Gabor coefficients transformed from 45 facial landmark points based on Active Appearance Model (Lucey et al., 2006), and classified AU combinations using a Sparse Representation (SR) classifier. Whitehill et al. (2011) detected 19 AUs by feeding 72 complex-valued Gabor filtered features to a separate linear SVM, and subsequently recognized six basic emotions using multivariate Logistic Regression (MLR) from the detected AUs. There are also some other facial action and emotion recognition approaches using static features that have been investigated, such as Local Binary Patterns (Shan et al., 2009) and Haar features (Whitehill & Omlin, 2006), etc.

The use of only static features, however, faces a drawback, i.e. the dynamic information of facial movements has been ignored and also the static features tend to vary a lot between different subjects (e.g. the shapes of eyes and the width of mouth). Thus it may lead to the inadequacy of generalization ability and efficiency. In order to address this issue, recently there have been great efforts made in capturing dynamic facial features or making use of temporal variation of facial measurements. For example, Besinger et al. (2010) tracked 26 facial feature points from five facial image regions (eyebrows, eyes and mouth), and used the displacements of them to recognize three basic emotions in image sequences. Wang & Lien (2009) employed 3D motion trajectories of 19 facial feature points as inputs to SVMs and HMMs for the recognition of seven AU combinations. Kotsia et al. (2008) recognized 17 AUs and seven facial expressions by the fusion of displacements of 104 Candide grid nodes and texture information features using SVMs and Median Radial Basis Functions (MRBFs) Neural Networks. Their average recognition rates for AUs and facial expressions were 92.1% and 92.3%, respectively. Tsalakanidou & Malassiotis (2010) proposed a rule-based real-time AU and emotion recognition system based on facial geometric, appearance, and surface curvature features extracted from 2D+3D images. Their results demonstrated good accuracy rates for the recognition of 11 selected AUs and four basic facial expressions. Srivastava & Roy (2009) used spatial displacements (or residues) of 3D facial points and SVM classifiers to recognize the six basic emotions, and demonstrated better recognition accuracies in comparison to the employment of pure static facial features (91.7% for dynamic features vs 78.3% for static features). Gong et al. (2009) employed shape deformation between an expressional 3D face and its corresponding reference (neutral) face to classify the six basic emotions using SVM classifiers. The estimation of the basic neutral facial shape was performed based on Karhunen-LoeveTransform (KLT), which is closely related to Principal Component Analysis. More recently, Valstar & Pantic (2012) used Gabor-feature-based boosted classifiers and particle filtering with factorized likelihoods to track 20 facial points through a sequence of images. These facial geometric points were then used as inputs to a hybrid classifier composed of Gentle Boost, SVMs, and hidden Markov models (HMMs) to recognize 22 AUs. They attained an average AU recognition rate of 72% when tested on spontaneous facial expression images.

More recently, Salahshoor & Faez (2012) proposed a novel dynamic mask to automatically segment the regions of face which were less sensitive to expressions and applied a modified nearest neighbor classifier for the recognition of the six basic emotions. Moreover, Ujir (2013) decomposed a face into six distinct regions and extracted their 3D facial surface normals instead of raw 3D points as the feature vectors. Then Support Vector Machines were employed to recognize facial expressions for the six regions independently. A weighted voting scheme was also applied to make the final classification.

Moreover, a variety of feature optimization methods has been successfully applied to facial expression recognition. For example, Tang & Huang (2008) performed automatic feature optimization by maximizing the average relative entropy of marginalized class-conditional feature distributions, and identified less than 30 discriminative features from the pool of all possible line segments between 83 landmarks. Their automatic feature selection achieved approximately 2% - 5% performance improvements for the recognition of the six basic emotions in comparison to their manually selected features. Soyel & Demirel (2009; 2010) adopted Principal Component Analysis to reduce the dimensionality of the raw feature set that consisted of distances between all possible pairs of 83 facial landmarks, and then applied Linear Discriminant Analysis (LDA) to find the optimal subset that preserved the most discriminant information. A two-stage probabilistic neural network was subsequently employed for the classification of seven facial expressions. Tekgüç et al. (2009) adopted the non-dominated Sorted Genetic Algorithm II for feature optimization, which is developed particularly to resolve problems of multi-objective aspects with more accuracy and higher convergence speed, and achieved an average recognition rate of 88.18% for the classification of neutral plus the six basic emotions. Pinto et al. (2011) employed a Sequential Forward Floating Selection (SFFS) algorithm to select the optimal subsets of features from different scales of 2D and 3D wavelet transform features extracted for seven expressions. Dornaika et al. (2011) evaluated the effect of applying Estimation Distribution Algorithm (EDA)-based feature optimization on a variety of machine learning algorithms for the recognition of different facial expressions from video sequences, including Naive Bayes, Bayesian Networks, Support Vector Machines, *K*-Nearest Neighbor, and Decision Trees. Their experimental results showed that the EDA-based feature selection significantly improved the recognition performance for all the above classifiers (3% - 18%). There are also other feature selection techniques that have been applied to facial expression analysis, such as GentleBoost (Sandbach et al., 2012), the Kullback-Leibler divergence measure (Tang & Huang, 2008), and the normalized cut-based filter (NCBF) algorithm (Sha et al., 2011).

Although the above systems showed noticeable improvements on recognition accuracy, many state-of-the-art AU and emotion recognition systems still suffer from the following problems. First of all, automatic AU intensity measurement posed great challenges to automatic recognition systems since the differences between some AUs' intensity levels could be subtle and subjective, and the physical cues of one AU might vary greatly when it occurs simultaneously with other AUs. Furthermore, FACS only defines a five point ordinal scale to describe the intensity of an AU. It does not define a quantifiable standard to measure the strength of corresponding facial changes. Hence, although there is substantial research concentrating on automatic AU recognition (e.g. Sorci & Thiran, 2010; Pantic & Patras, 2006; Tong et al., 2007; Li et al., 2013), the companion problem of accurately estimating the AU intensity levels has not been much investigated. There were only limited applications in the literature on AU intensity estimation. For instance, Kaltwang et al. (2012) realized continuous AU intensity estimation based on facial landmarks and appearance features by using a set of independent regression functions, but the work only focused on 11 specified AUs that were closely related to the recognition of shoulder pain facial expressions. Bartlett et al. (2006) found that in AU classification tasks, distances between samples to SVM separating hyperplanes were correlated with AU intensities. Based on this finding, Savran et al. (2012) realized intensity estimation of 25 AUs from still images on both 2D and 3D modalities using appearance features and regression based methods. They claimed that the proposed approach for AU intensity estimation performed better than other state-of-the-art methods (average correlations of 54.3% for lower face AUs and 74.4% for upper face AUs).

Furthermore, in contrast to AU detection, robust facial emotion recognition using AU intensities is still largely unexplored. Current research mainly focused on rule-based and statistical-based methods. For example, Valstar & Pantic (2006) explored both a formulated rule-based method and an Artificial Neural Network (ANN) based method to predict emotions from AUs. However, their recognition accuracies still required further improvements. It could be attributed to the fact that the former, i.e. the rule-based reasoning, was not robust enough to deal with noises and errors, while the latter, i.e. directly using machine learning techniques, relied on extensive training data to accommodate possible AU combinations for each emotion category. Chang et al. (2009) proposed a hidden conditional random fields (HCRFs) based method to map various combinations of 15 most frequently occurring AUs to underlying emotions, but extensive annotation work was required prior to mapping.

Finally, although equipped with appropriate domain knowledge, manual feature selection is often time consuming and requires an endless trial-and-error process, there are also extensive optimization algorithms and boosting techniques devoted to automatic feature selection and feature dimensionality reduction including Principle Component Analysis (PCA), Fisher Linear Discriminant (FLD), genetic and evolutionary algorithms, and AdaBoost. PCA has been widely used for feature selection for face and facial expression recognition for decades (Jeong et al., 2009). According to Swets & Weng (1996), PCA derives most expressive features but may not embed sufficient discriminating power. FLD is another commonly used feature reduction technique which is claimed to provide comparatively more class separability by maximizing the mean between classes and minimizing the variation within a class

(Chavan & Kulkarni, 2013; Gu *et al.*, 2012). However, it requires a wide coverage of face/class variations at the training stage in order to get more superior recognition performance.

Thus, we aim to overcome these challenges discussed above, and develop a practical, robust and person-independent solution for real-time Action Unit intensity estimation and emotion recognition. We employ automatic-selected motion-based facial features with a strong psychological background to estimate the intensities of the 16 AUs closely associated with the expression of the six basic emotions. Subsequently, the 16 AUs are ranked for each emotion according to their discriminative power. The derived intensities of the most discriminative AU combinations are then respectively employed as inputs to a set of six novel ensemble classifiers to robustly recognize the six basic emotions regardless of errors and noises involved in the input AU intensities. The proposed ensemble classifiers also have great capability to identify newly arrived unseen novel emotions. The details of the proposed facial expression recognition system are presented in Chapter 3.

2.5 Automatic emotion recognition from bodily expressions

Having discussed the huge progress in emotion recognition from facial expression, we now present the state-of-the-art research of automatic emotion recognition from bodily expressions. As mentioned earlier, most efforts conducted so far on automatic emotion recognition have concentrated on the facial modality, only until recently there have been some automatic systems that are able to detect emotions based on bodily behaviors.

Most recent automatic emotion recognition research makes use of either body posture or movement as the source of affective information. For example, many early developments have focused on recognizing emotions from expressive dance (e.g. Camurri et al., 2003; Camurri et al., 2004; Park et al., 2004; Kamisato et al., 2004).

Camurri and colleagues (Camurri et al., 2003; Camurri et al., 2004) extracted dynamic motion cues from dancers' body movements to differentiate between basic emotions. They found significant relations existed between certain emotion categories (e.g. happy and exciting) and key motion qualities (e.g., body contraction index, fluency of motion, and time duration). Kapur et al. (2005) used a 3D motion capture system to record dancers' posed bodily movements. The participants were instructed to freely enact four emotional states: sadness, joy, anger, and fear. The experimental results showed that the Support Vector Machine based classification was able to achieve an average classification accuracy of 91.8%, while human observers achieved an average of 93% on the same data.

These results demonstrated that body movements are an effective channel for automatic emotion recognition in either acted or expressive dance scenarios. However, in many cases, subjects are instructed to perform certain emotions, thus the recorded movements are exaggerated and purposely geared toward emotional expressions whereas bodily expressions in everyday scenarios are more subtle and thus inevitably pose more challenges automatic emotion decoding systems. Castellano et al. (2007) proposed a bodily expression recognition system which focused on non-propositional movement qualities of arms (e.g. velocity, amplitude and fluidity of movement) rather than static gesture shapes. An average recognition rate of 61% was achieved by their Bayesian Network-based classifier for the recognition of anger, joy, pleasure and sadness. Bianchi-Berthouze & Kleinsmith (2003) proposed a categorical approach to recognize three discrete emotions (anger, happiness and sadness) from 138 acted posture images. By using low-level descriptions of body postures, they obtained an overall classification rate around 95%. Their most recent follow-up work (Kleinsmith et al., 2011) considered non-acted postures and more subtle bodily expressions in the gaming scenarios. An average accuracy rate of 59.22% was achieved for the classification of the following four emotion categories: 'concentrating', 'defeated', 'frustrated' and 'triumphant'. These results seemed considerably worse than typically quoted rates achieved from the acted and dance-based systems discussed above. It also indicates the challenges of detecting discrete emotions from subtle bodily expressions in real-life scenarios.

More recently, the use of dimensional representation of emotions has shown great potential in automatic emotion recognition from bodily expressions. Many existing efforts on dimensional emotion recognition have tended to quantize the dimensional values into discrete levels, e.g. the work of Fragopanagos & Taylor (2005) which reduced the dimensional value prediction problem to a four-class classification problem, i.e. classification into one quadrant of a 2D Valence-Arousal space (positive vs. negative; active vs. passive). There are also other more comprehensive systems (e.g. Kleinsmith & Bianchi-Berthouze, 2007; Karg et al., 2010; Kleinsmith et al., 2011) that attempted to quantize the continuous range of each dimension into certain levels (e.g. 3-7 point Likert scales). On the contrary, Kleinsmith & Bianchi-Berthouze (2013) indicated that a continuous representation of affective dimensions may provide a more accurate and generic measurement of users' emotional states. However, relatively less effort has been made to interpret emotions in continuous dimensions. Gunes and Pantic (2010) focused on dimensional emotion recognition from head gestures in spontaneous conversations. They employed features of head motion, direction, and the occurrences of head nod and shake to estimate continuous levels of the arousal, valence, intensity and expectation dimensions, and achieved an average Mean Squared Error (MSE) of 0.102 using Support Vector Regression (SVR). Nicolaou et al. (2011) proposed a multimodal system for continuous and dimensional emotion prediction of a speaker. They employed various modalities including facial expression, shoulder gesture and audio cues to continuously track the levels of the valence and arousal dimensions by using SVR and Long-Short Term memory (LSTM) regression.

Although dimensional affect recognition from bodily expressions has drawn increasing research attention, most existing systems either focused on specified parts of the body (e.g. head gestures in the work of Gunes and Pantic (2010)), or only considered either static body form or dynamic motion information as the source of interpretation. For example, Kleinsmith et al. (2011) focused on static posture features whereas Savva et al. (2012) utilized purely body motion features for emotion interpretation. There are only a few systems that have fully considered both form and motion information from whole-body expressions for emotion recognition. One notable milestone is the recent work by Metallinou et al. (2013), who addressed the problem of tracking continuous levels of valance, arousal and dominance by using full-body language features in inter-personal interactions. In their work, a 3D Motion Capture system with 12 Vicon MoCap cameras was employed to capture participants' whole-body expressions. Both body posture (e.g. head rotation, hand position and body leaning angle) and movement (e.g. velocity of arm/foot) features were extracted accordingly, and then inputted to a Gaussian Mixture Model (GMM) to estimate the underlying emotional state. They also produced a statistical analysis of each single bodily feature in order to select a subset of de-correlated informative features for each affective dimension. Promising results were obtained for the tracking of the arousal and dominance dimensions (median correlation = 0. 584 and 0.37, respectively). However, significantly lower performance was observed for the valence dimension (median correlation = 0.225). This may be attributed to inadequate features employed for the reflection of valence. For example, the dynamic features they employed were only concerned with velocity, however, other informative types of features such as acceleration, frequency and amplitude were ignored. Moreover, although a statistical feature selection was performed for each feature, both body form and motion features were mixed indiscriminately for the prediction of each affective dimension, which is controversial to the psychological findings discussed earlier (i.e. static form and dynamic motion features could contribute distinctively to different affective dimensions), thus may also lead to performance drop.

Moreover, we briefly presented some techniques that have been successfully applied for bodily feature selection. For example, De Silva & Bianchi-Berthouze (2004) employed Discriminant Analysis to measure the saliency of the proposed 24 emotional posture features and identified a number of feature subsets which can explain the separation between different emotions. Kleinsmith & Bianchi-Berthouze (2007) applied non-linear mixture discriminant analysis (MDA) to select the most discriminating features from 24 low-level posture features for the discrimination between pairs of affective dimension levels (e.g., low vs. high, etc.). The MDA conducted an iterative process to create different models based on linear combinations of the most discriminating features, so that it was able to ascertain the optimal feature sets that led to the best performance. Bernhardt (2010) proposed a merit function to evaluate both of the class-feature correlations and the inter-feature correlations of a certain feature subset based on the information-theoretic concept of information gain, and then used this function as a heuristic to direct a Hill-climbing algorithm to identify an optimal feature subset that was highly correlated with the emotional classes, but uncorrelated with each other.

Compared to the developments discussed above, this research presents an effective solution for real-time and dimensional affect recognition based on whole-body expressions. We comprehensively consider both static posture (e.g. distances, body leans and joint angles) and dynamic motion (e.g. velocity, acceleration and amplitude) features to draw a more comprehensive representation of whole-body behaviors. We then employ the GA for automatic feature optimization and selection. For robust prediction of users' continuous affective dimensions, we propose an adaptive ensemble regression model which also has great capability to adapt to new bodily expressions and deal with data stream regression. We also examine the roles of both body form and motion features in predicting each of the affective dimensions (i.e. arousal and valence), and then identify their optimal combination tailored to each affective dimension. In additional, as pointed out by Kleinsmith & Bianchi-Berthouze (2013) and Metallinou et al. (2013), a challenging issue of emotional data annotation is that a high level of disagreement may arise when building the ground truth, especially for continuous and dimensional annotation tasks. Therefore, we also propose a novel annotation method using both the correlation between different annotators and the personal bias metrics to effectively establish the ground truth for evaluation. The details of the proposed dimensional affect interpretation are presented in Chapter 4.

Chapter 3 Facial action unit intensity estimation and expression recognition

In this chapter, we present the adaptive facial expression recognition system in detail. For a more comprehensive understanding, we first of all provide an overall description of the proposed system, which is composed of: facial geometric data tracking, mRMR-based feature selection, Action Unit intensity estimation using Neural Networks (NN) and Support Vector Regressors (SVR) and emotion recognition with adaptive ensemble classifiers. Figure 3-1 illustrates the system's overall architecture and dataflow.

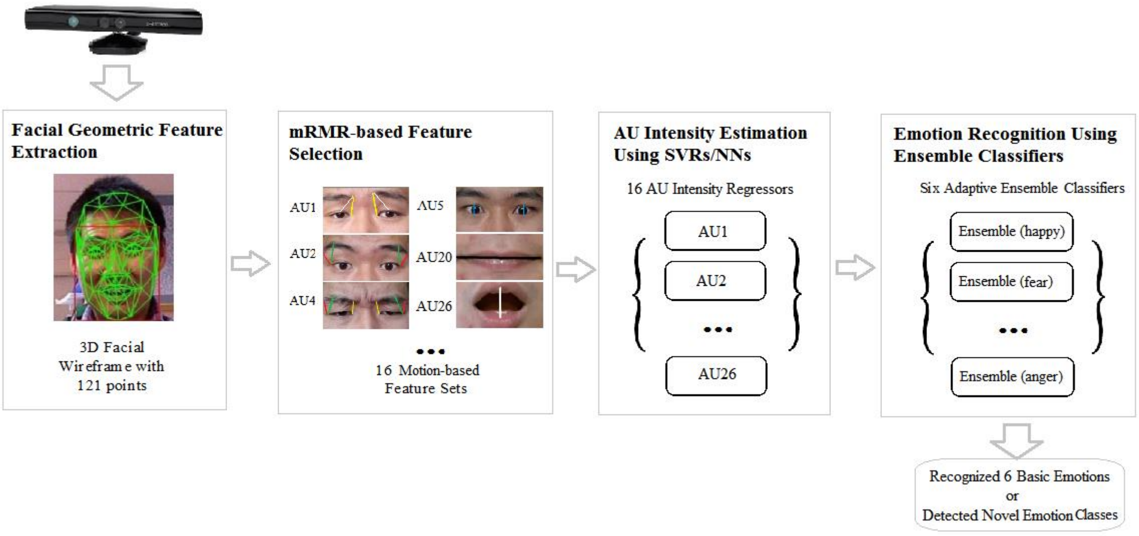

Facial Geometric Feature Extraction
3D Facial Wireframe with 121 points
mRMR-based Feature Selection
AU1
AU2
AU4
AU5
AU20
AU26
16 Motion-based Feature Sets
AU Intensity Estimation Using SVRs/NNs
16 AU Intensity Regressors
AU1
AU2
AU26
Emotion Recognition Using Ensemble Classifiers
Six Adaptive Ensemble Classifiers
Ensemble (happy)
Ensemble (fear)
Ensemble (anger)
Recognized 6 Basic Emotions
or
Detected Novel Emotion Classes

Figure 3-1 System architecture and data processing pipeline

The main processing of the facial emotion recognition system includes the following.

1. The real-time facial geometric data tracking is implemented based on a Microsoft Kinect sensor (Webb & Ashley, 2012) and a variant of Candide-3 model (Ahlberg, 2001). The Kinect's facial analysis API is able to localize a total of 121 3D facial landmarks and perform continuous tracking at a frame rate of 25~30 fps.

2. We extract motion-based facial features for AU intensity estimation, which are calculated based on facial wireframe node displacements. The motion-based features are caused by underlying facial muscle movements, and thus are relatively

universal and subject-independent, and contain comparatively richer emotional information compared to static features.

3. We then apply both manual and mRMR based automatic feature selection methods to select 16 sets of informative features from the complete pool of candidate features for the regression of 16 diagnostic AUs. The feature sets selected are respectively employed as inputs to 16 AU intensity estimators, with each estimator dedicated to each AU. We employ Neural Networks and Support Vector Regressors for AU intensity estimation.

4. For robust emotion recognition, the 16 diagnostic AUs are first ranked and filtered according to the AU-Emotion relationships with intention to identify the most discriminative AU combinations for each emotion category. We then propose six novel adaptive ensemble models for robust classification of the six basic emotions and novel emotion detection, with each ensemble dedicated to each emotion category.

The remainder of this chapter is organized as follows: Section 3.1 discusses the method employed for raw facial geometric feature points tracking. Section 3.2 presents the detailed procedures of AU intensity estimation, including motion-based feature extraction, both manual and mRMR based feature selection, and intensity estimation using NN and SVR. In Section 3.3, we discuss the proposed adaptive ensemble scheme for the challenging task of emotion recognition and novel unseen emotion detection using selected AU intensities. The experiments and both on-line and off-line evaluations for AU intensity estimation and emotion recognition are discussed in Section 3.4. We draw conclusions for this chapter in Section 3.5.

3.1 Facial geometric information tracking

Regarding to 3D facial geometric feature extraction, a number of well-known methods have been examined, such as the Kanade-Lucas-Tomasi (KLT) tracker

(Bouguet, 1999) and the Vukadinovic-Pantic facial point detector (Vukadinovic & Pantic, 2005). Both of them are able to generate good tracking results with static input images, but limitations rise up when dealing with real-time 3D streams. In our system, the 3D face geometric data are acquired through a Kinect and its embedded face tracking engine (Webb & Ashley, 2012). The Kinect is an effective research tool that physically integrates a color camera with up to 1280 x 960 resolutions, a depth-sensing camera with up to 640 x 480 resolutions, and an array of four microphones. It provides efficient real-time 3D tracking capabilities in a relatively inexpensive package.

When emotions are being expressed by a subject, the facial elements change their shapes and positions accordingly. These geometric changes caused by facial muscles contain rich motion-based facial features. Once completing parameter adjustments and successfully detecting a user's face, the Kinect face tracking engine performs fitting and subsequently tracks a 3D variant of the Candide-3 model with 121 grid nodes. The facial tracking algorithm makes use of both color and depth image data streams to reconstruct salient facial models, enabling better robustness against variations in illumination, scaling, skin color and especially head poses. In good lighting conditions, it is able to track a face reliably when the user's head pitch, roll and yaw are respectively less than 10, 45 and 30 degrees (Webb & Ashley, 2012).

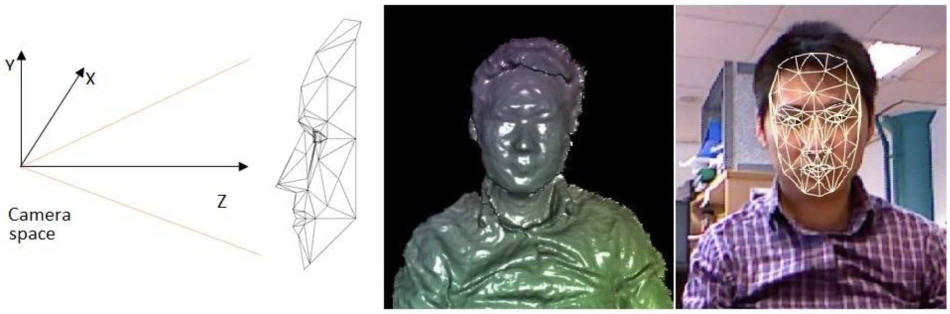


Figure 3-2 The Kinect 3D coordinate system (left), 3D surface reconstruction with depth data (middle) and a tracked 3D facial wireframe (right)

The tracked facial wireframe is able to automatically fit to the detected face in the Kinect 3D coordinate space and evolves through the video sequence (see Figure 3-2). It is able to reach up to 30 fps on i7 quad-core CPUs with 8GB RAM. If required, the loss or error of tracked wireframes could be handled by a model deformation algorithm, which is able to add mesh fitting at the intermediate steps of tracking. Such a procedure increases robustness against node losses and ensures tracking effectiveness. An essential normalization procedure is also performed afterwards, where the information of head orientation and distance to the sensor is employed to adjust the tracked facial grid model. Figure 3-3 shows a neutral state plus facial expressions for the six basic emotions associated with generated corresponding 3D facial wireframes.

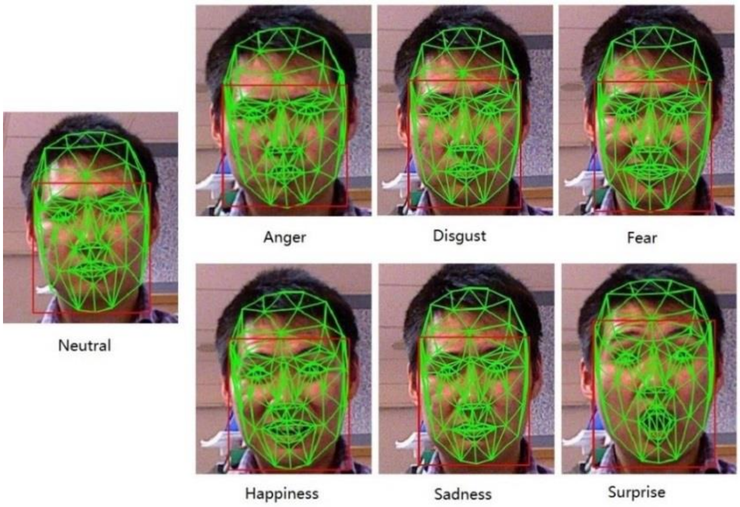


Figure 3-3 Examples of tracked 3D facial wireframes for each expression (The green lines represent facial wireframes, while the red rectangles indicate detected facial areas)

3.2 Facial action unit intensity estimation

In the literature, most recent research work employed either image driven or prior model-based methods for automatic AU recognition. The former (e.g. Chang et al., 2004) performed recognition based on static image data directly while the latter was developed to extract the relationships and spatial-temporal information of AUs using prior models (e.g. Tong et al., 2010; Valstar & Pantic, 2007). However both required a considerable amount of reliable training data, which sometimes could be difficult and expensive to acquire. More importantly, generalizing a model trained on one database to other databases could still be a challenging issue, especially for real-life applications

(Li et al., 2013; Torralba & Efros, 2011). In order to overcome these challenges, we propose and employ motion-based facial features, which are supported by psychological studies and facial anatomy, and thus are more pertinent for AU intensity estimation. The 16 AUs we focus on in this research are AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller), AU13 (Cheek Puffer), AU15 (Lip Corner Depressor), AU17 (Chin Raiser), AU18 (Lip Puckerer), AU20 (Lip Stretcher), AU23 (Lip Tightner), AU24 (Lip Pressor), AU26 (Jaw Drop) and AU27 (Mouth Stretch). Compared to existing research on AU detection, our work has the following two advancements:

1. We propose dynamic motion-based facial features (e.g. the elongation of mouth) for AU intensity estimation, which can be measured through the displacement of facial points between natural and expressive frames. As discussed earlier, such features are caused by underlying facial muscle movements, and thus are relatively universal and subject-independent.
2. We apply both manual and automatic methods to select a unique subset of informative features for each AU respectively. The manual feature selection is guided by FACS domain knowledge, while the automatic feature selection is performed based on mRMR based optimization (Peng et al., 2005). Their performance and comparison are presented in Section 3.4.2.

3.2.1 Extraction of motion-based facial features

As a part of MPEG-4 FBA [ISO14496] International Standard, the MPEG-4 face animation framework (Pandzic & Forchheimer, 2012) is designed to deal with face animation applications, including reproduction of facial shape, texture, subtle expressions, as well as speech pronunciation. MPEG-4 defines 84 facial feature points to best reflect the facial anatomy and movement mechanics, which are learned from subtle facial actions and are closely related to muscle actions, as illustrated in Figure

3-4 (Pandzic & Forchheimer, 2012). Based on this standard, we derive a series of 3D distance features between key facial points, and then use dynamic changes of these distances for AU intensity estimation.

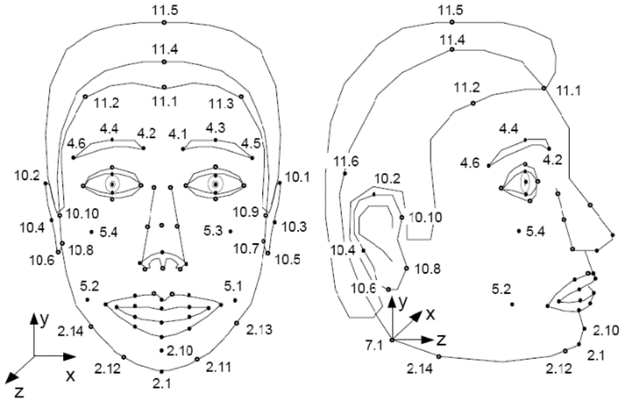


Figure 3-4 Facial feature points defined in MPEG-4 (Pandzic & Forchheimer, 2012)

When reliably detecting a user’s face, the face tracking component continuously outputs a sequence of normalized 3D facial wireframes (compatible with MPEG-4 standard) in a real-world 3D coordinate system. Each wireframe consists of 121 grid nodes, including 16 nodes for eyes (i.e. 8 nodes for each eye contour), 20 nodes for eyebrows (i.e. 10 for each eyebrow), 12 nodes for the upper lip, 16 nodes for the lower lip, 16 for the nose, and others for making up the rest of the mesh model. The tracking process of 3D geometrical feature points is also robust to head rotations up to 10, 45 and 30 degrees in pitch, roll and yaw as discussed above.

We first of all acquire reference measurements of the neutral facial expression of each subject. Rather than requiring subjects to deliberately pose an initial calibration expression of the neutral state (which is often unreliable), we record the first 50-100 frames (typically 2-4 seconds, when subjects are naturally in their neutral states), and then compute the median data of these neutral frames to form a set of reference measurement vectors $\{R_i\}$ for the representation of neutral faces.

The motion-based facial features can be then calculated through facial point displacements between natural and expressive frames. Equations (3-1) and (3-2) define the calculation of any motion-based facial feature in the 3D Euclidean space.

$$d_{i,j} = \sqrt[2]{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (3\text{-}1)$$

$$\Delta d = d_{i,j}(expressive) - d_{i,j}(neutral) \quad (3\text{-}2)$$

In Equation (3-1), $d_{i,j}$ is the distance between *node*$_i$ (i.e. a 3D facial feature point i) and *node*$_j$ (i.e. a 3D facial point j) among the generated 121 3D facial wireframe nodes, and in Equation (3-2), Δd defines the change of distance feature $d_{i,j}$ between the reference (neutral) frame and any expressive frame. Such distance features are computed based on a real-world 3D coordinate system. As discussed before, the facial tracking engine of the Kinect is able to perform face fitting with high accuracy and is also able to identify the distances of different facial regions to the sensor using depth images obtained from its depth camera to deal with facial geometric feature tracking with head rotations and movements. Thus, our facial tracking component developed based on such a platform is capable of providing robust fitting and 3D geometric feature extraction to deal with head pose variations and movements in real-life applications.

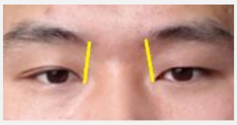
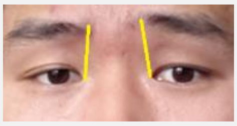
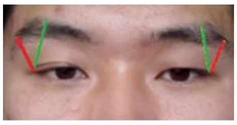
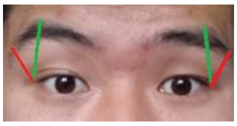
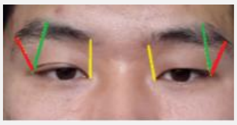
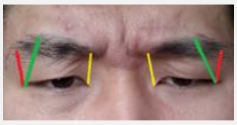
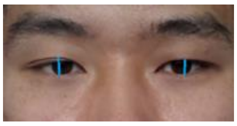
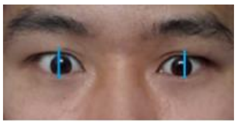
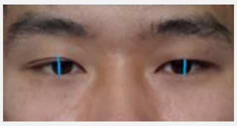
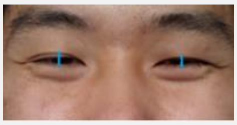
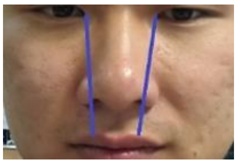
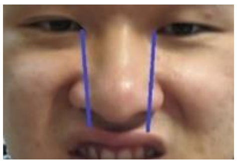
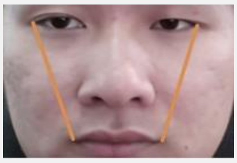
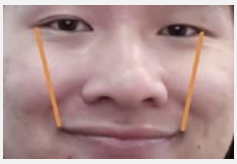
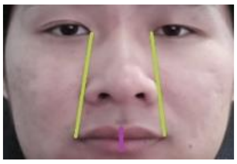
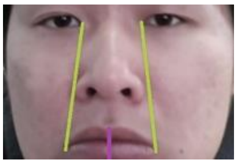
However, n number of facial feature points will result in a large number of C_n^2 unique distance features (e.g. 121 facial points will produce C_{121}^2= 7260 distance features). Intuitively, not all of the distance features are informative for the detection of a specific AU. Thus, rather than applying the distance features between entire facial points for all AUs without distinction (e.g. Kotsia et al., 2008), we next step focus on generating a subset of informative discriminating features from the candidate feature pool for each AU respectively, which may lead to optimized performance.

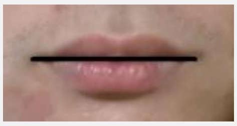
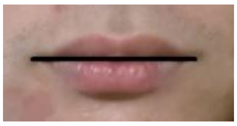
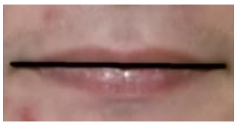
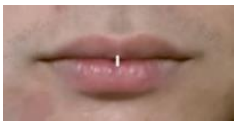
3.2.2 Feature selection for AU intensity estimation

3.2.2.1 Manual feature selection

In typical manual feature selection, the features are derived based on sufficient domain knowledge. We extract a total of 24 representative facial motion-based features (i.e. Δd distance changes) using 22 key facial feature points out of the whole 121 points, as illustrated in Table 3-1. According to Ekman & Friesen (1983) and Ekman et al. (2002), these features are believed to play an important role in determining the level of AU intensities. As shown in Table 3-1, each AU is associated with a subset of features composed of only a small number of relevant features (typically 2 to 6 dimensions). Such features are derived according to FACS domain knowledge, and we especially focus on analyzing the movement of facial muscles underlying each AU for subsequent AU intensity estimation.

Table 3-1 Examples of manually selected features and measurements represented by lines of different colors

AU	***Measurement Nodes***	***Distance Features (Neutral)***	***Distance Features (Expressive)***
AU1 Inner Brow Raiser	Inner eyebrow corner, inner eye corner		
AU2 outer Brow Raiser	Outer eyebrow corner, Outer eye corner, Middle top of eyebrow		
AU4 Brow lower	Eyebrow corners, Inner/outer eye corner, Middle top of eyebrow		
AU5 Upper Lid Raiser	Middle eyelid top, Middle eyelid bottom		
AU6 Cheek Raiser	Middle eyelid top, Middle eyelid bottom		
AU10 Upper lip Raiser	Inner eye corner, Right/left top of upper lip		
AU12 Lip Corner Puller	Outer eye corner, Right/left mouth corner		
AU15 Lip Corner Depressor	Inner eye corner, Mouth corner, Top/bottom of lips		

AU18 Lip Pucker	Right/left mouth corner		
AU20 lip Stretcher	Right/left mouth corner		
AU23 Lip Tightner	Right/left top/bottom of lips		
AU26 jaw Drop	Middle bottom/top of lips		

Moreover, for a deeper understanding, we provide two examples for manual feature selection in the following. For example, when AU1 (Inner Brow Raiser) is occurring for a specific facial emotion expression, the inner portion of the eyebrows is pulled upwards by muscle 1 (see Figure 3-5) (Ekman et al., 2002). This causes an inevitable increase in the distance between inner eyebrow corner and inner eye corner. Thus, the distance variation Δd between the neutral and this expressive frame may contribute to the estimation of the occurrence and intensity of AU1.

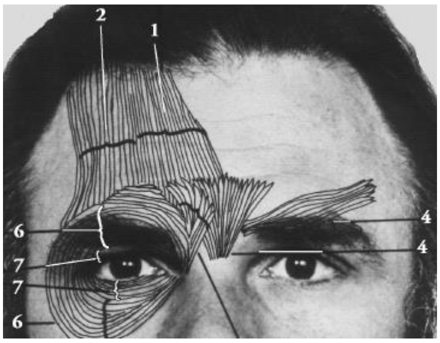


Figure 3-5 Muscles associated with upper facial Action Units (Ekman et al., 2002)

Furthermore, the following indicates a slightly more complicated example. AU12 (Lip Corner Puller) and AU13 (Sharp Lip Puller) are often accompanied by a smile or a joyful facial expression. These AUs are caused by pulling the corners of the lips back and upwards to form a ︶ shape of the mouth. But it is unlikely that we can directly use some intuitive distance features, such as the elongation of the mouth, to distinguish these AUs (although the mouth is indeed elongated). The reason is that there are other

AUs that can also cause mouth elongation, such as AU20 (Lip Stretcher). Thus the extraction of distance features becomes challenging. However by analyzing these facial movements from the perspective of anatomy, we can see there are two underlying muscles related to these AUs - *Zygomaticus Major* [12] and *Caninus* [13], as shown in Figure 3-6 (Ekman et al., 2002). Both originate on the upper cheek bone and attach with the corner (angle) of the lips. When contracted, they will pull the corners of the mouth naturally up towards the upper cheek. Thus, the distances between mouth corners and outer eye corners are reduced synchronically. Therefore, we select the outer eye corners as the reference points for AU12 or AU13, because their positions are relative fixed and can be tracked reliably.

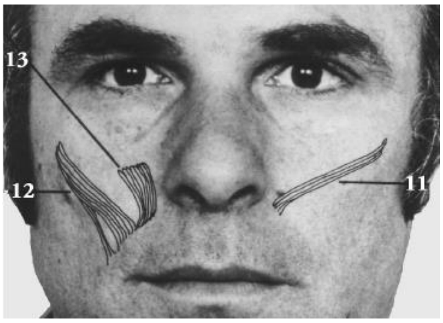


Figure 3-6 Locations of muscles underlying lower facial oblique Action Units (Ekman et al., 2002)

Note that, in this research, Δd can be either positive or negative. For instance, AU1 (Inner Brow Raiser) may cause a positive Δd which means an increase in distance between inner eye corners and eyebrow corners. When Δd becomes negative, it indicates the eyebrow is lowered, which means AU4 (Brow Lowerer) occurs. Table 3-1 summarizes some AUs and their corresponding manually selected features, and gives a clear illustration on how they change synchronically with the occurrence of each AU (for clarity, all samples showed in Table 3-1 are in 2D although in the real system, 3D facial points are extracted as discussed in Section 3.1). The above FACS domain knowledge-based manual feature selection provides an efficient and robust approach against facial shape variations of different subjects.

3.2.2.2 Automatic feature selection based on mRMR

As the most common form of evolutionary optimization, conventional genetic algorithms evolve a large population of candidate solutions by mimicking the process of natural selection (Sikora & Piramuthu, 2007). Other commonly used evolutionary algorithms include Particle Swarm Optimization (Wang et al., 2007) and Genetic Programming (Davis et al., 2006), etc. However, applying such algorithms in a large search space (e.g. thousands of dimensions) tend to be very computationally exhaustive and time consuming. Furthermore, inappropriate parameter configuration may easily lead to premature convergence to a local extremum. On the contrary, mutual information (MI) is information based feature selection that is not limited to linear dependencies, and is able to maximize information in a class. Research on the performance improvement of MI has brought to the development of minimal-redundancy-maximal-relevance criterion, which is a variant of MI. In this research, since a large proportion of the raw facial distance features could be less informative or considerably redundant with each other, it is reasonable to apply information theory based methods for automatic feature selection, which could well reflect relevance between features and outputs and within features comprehensively. Moreover, such methods also have relatively lower computational complexity and better generalization of the selected features on different classifiers. Thus, we are motivated by mRMR to propose an attractive alternative for automatic feature selection.

Tang and Huang (2008) proposed a novel method based on maximizing the average relative entropy of marginalized class-conditional feature distributions, and successfully applied it to 3D facial distance feature selection tasks. Their automatically selected features achieved higher recognition accuracies than their manually devised features for the six basic emotions (about 2% - 5% improvements). However, their method is difficult to directly apply to regression problems as the lack of effective relevant calculation method for continuous values. Thus, we introduce a modified mRMR-based feature selection method to deal with the case where both features and outputs are continuous data.

We introduce the mRMR optimization algorithm in the following. mRMR is introduced by Peng et al. (2005) and aims to minimize the mutual information between the selected features (i.e. redundancy), and to maximize the mutual information between the selected features and the desired output (i.e. relevance). Let x_i denote a feature and $S_M = \{x_i\}_{i=1}^{M}$ be an instance consisting of *M* features. *I* denotes the mutual information with *y* indicating the desired output, and $p(x_i)$, $p(y)$, $p(x_i, x_j)$, and $p(x_i, y)$ representing the probabilistic density functions. Then the traditional mRMR measure can be described as follows:

$$mRMR(i) = I(x_i, y) - \frac{1}{M-1}\sum_{x_j \in S_M, j \neq i} I\left(x_i, x_j\right) \qquad (3\text{-}3)$$

where

$$I(x_i, y) = \sum_{x_i \in X_i} \sum_{y \in Y} p(x_i, y) \log\left(\frac{p(x_i, y)}{p(x_i)p(y)}\right) \qquad (3\text{-}4)$$

Since both the features and AU intensities in our system are continuous values, their mutual information is often hard to compute. I.e. it is difficult to compute the integral in the continuous space using a relatively limited number of samples. One solution is to perform a uniform data discretization processing in advance of the estimation of the mutual information value. However, this may lead to considerable information loss.

An alternative solution is to use linear correlations to approximate the mutual information, as suggested by Metallinou et al. (2013). By replacing the traditional mutual information metric with the Pearson correlation coefficient (CORR) (David, 2009), the mRMR measure can be well adapted to continuous values. The CORR represents the linear relationship between a pair of values, defined as follows:

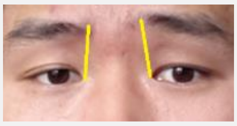
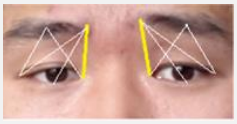
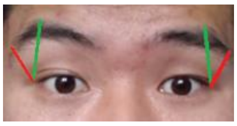
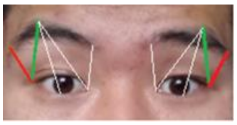
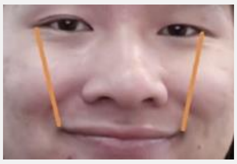
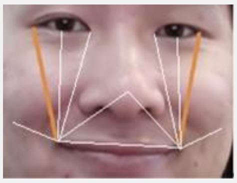
$$CORR(x, y) = \frac{COV\{x, y\}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad (3\text{-}5)$$

where COV stands for the covariance, and σ stands for the standard deviation, while “¯” symbolizes the mean.

$$mRMR(i) = |CORR(x_i, y)| - \frac{1}{M-1} \sum_{x_j \in S_M, j \neq i} \left|CORR(x_i, x_j)\right| \tag{3-6}$$

Then we perform a ranking of features according to their mRMR values. A higher value is preferred and it indicates that a specific feature contains more discriminating information, i.e. it has higher correlation with the desired output (e.g. 0.7) and lower correlation with other features (e.g. 0.3). We try different numbers of top ranking features as the inputs for AU intensity estimation, and those leading to the best performance are determined as the optimal features for each AU, as illustrated in Table 3-2. Evaluation results indicate that the proposed mRMR-based feature selection yields comparable results for AU intensity estimation when compared with the manual feature selection process.

Table 3-2 Comparison of manually selected features with those automatically selected by mRMR

AU	***Manually Selected Features***	***Automatically Selected Features***	***Dimensions of Automatically Selected Features***
AU1 Inner Brow Raiser			10
AU2 outer Brow Raiser			10
AU12 Lip Corner Puller			11

3.2.3 AU intensity estimation using selected features

For the task of automatic AU intensity estimation, we notice the following challenges. First, because of individual differences among subjects, overlapping between intensity levels (Savran et al., 2012) and annotators’ subjectivity are inevitable.

Second, the relationship between AU intensity levels and the scale of evidence might be nonlinear. To solve these problems, we employ two widely accepted algorithms, feedforward Neural Network with Backpropagation (Hecht-Nielsen, 1989) and Support Vector Regression (Vapnik, 1995) for AU intensity estimation, because of their effective handling of data comprising noises and non-linear relations. We also aim to examine the effectiveness of the mRMR based optimization in comparison to the manual feature selection, and to determine whether the features selected by mRMR are effective enough for discriminating between different levels of AU intensities.

3.2.3.1 Feedforward Neural Networks for regression

A feedforward Neural Network (BPNN) has the following two characteristics well suitable to our application:

1. It is robust to the noise and errors involved in training data, which may be inevitable in many supervised applications as mentioned above (Mitchell & Hill, 1997).
2. It needs some training costs, which depend heavily on the sample size, the dimensions of the training data, and the accuracy requirements. Once the model trained, however, it is extremely fast to be applied to the subsequent test instances. This would be beneficial to our real-time application.

A continuous value ranging from 0 to 1 is used as the single output to cover the whole interval of AU intensity levels (‘0’ represents absence with ‘1’ indicating maximum AU intensity). In this way, we can preserve sufficient AU intensity information for subsequent emotion recognition. Thus, we have the training data format as follows:

$$dataset_n = \{\Delta d_1, \Delta d_2, \Delta d_3, \dots, \Delta d_i, I\}$$

where the inputs Δd are the informative motion-based facial features for each AU selected by either the manual process or the mRMR-based optimization, and the output, *I*, is the ground truth intensity of that AU. Both the training and testing datasets are scaled using the same procedure before applied into Neural Networks in order to achieve the best performance (i.e. linearly scaling each attribute to the range of [-1; +1] or [0; 1]). We implement 16 single-hidden layer feedforward Neural Networks. Each of them has an input layer, a hidden layer, and an output layer (as shown in Figure 3-7). Each layer contains a number of nodes, which are interconnected with adjacent layers. Each node is a simple processing element that responds to the weighted inputs received from the preceding layer. The number of the nodes in hidden layer is set to 3-6 based on the complexity of the input layer.

The feedforward Neural Networks are trained by Backpropagation algorithm (BPNN) (Hecht-Nielsen, 1989). The Backpropagation iteratively adjusts the weights between the nodes in response to the errors until some targeted minimal error is achieved between the actual and target output values. The detailed method is shown in Algorithm 3-1. We also adjust the learning rate, the momentum and the termination error parameters to moderate values (e.g. respectively 0.1, 0.8, and 0.01), so that it is able to best achieve a balance between accuracy, speed and generalization performance.

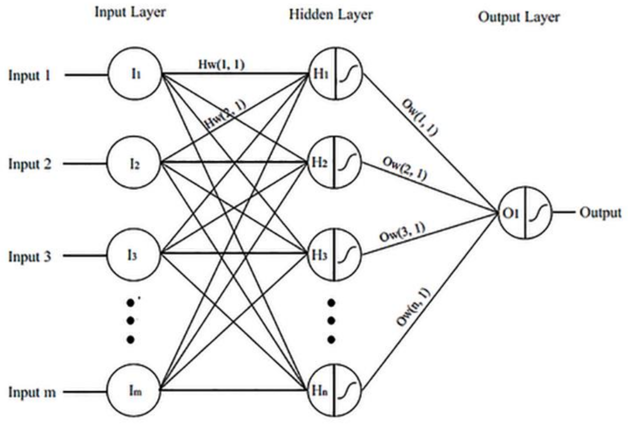
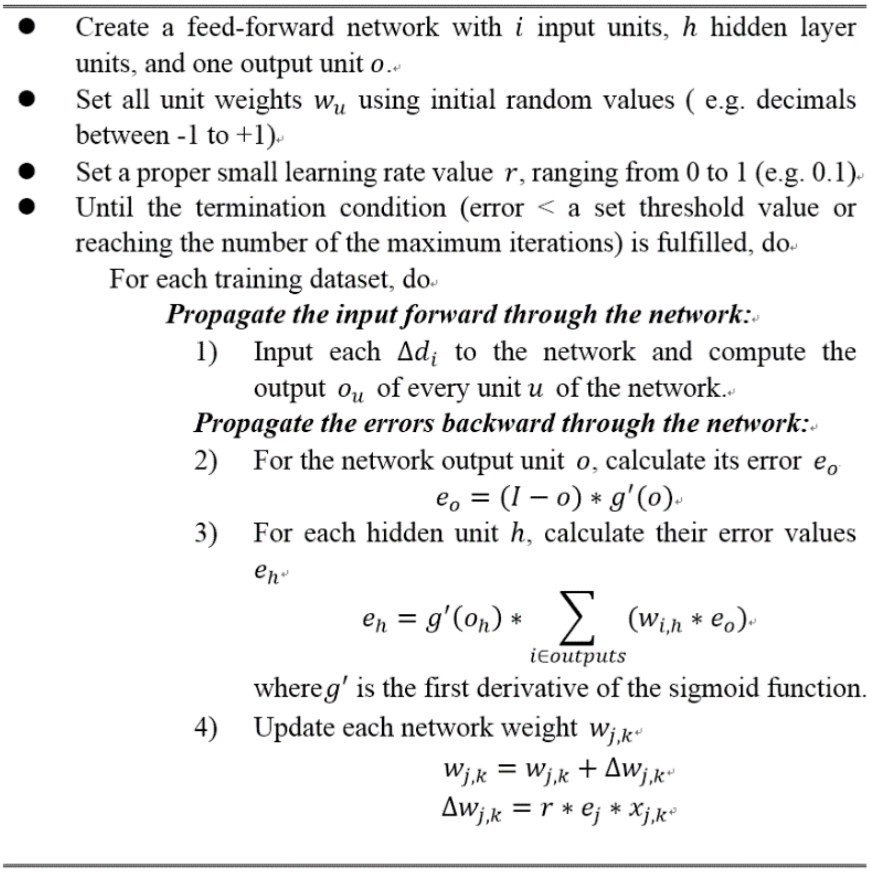


Figure 3-7 A sample topology of a single-hidden layer feedforward neural network

Algorithm 3-1 The training algorithm of the Neural Networks for AU intensity estimation (Hecht-Nielsen, 1989)



```
● Create a feed-forward network with $i$ input units, $h$ hidden layer units, and one output unit $o$.
● Set all unit weights $w_u$ using initial random values ( e.g. decimals between -1 to +1).
● Set a proper small learning rate value $r$, ranging from 0 to 1 (e.g. 0.1).
● Until the termination condition (error < a set threshold value or reaching the number of the maximum iterations) is fulfilled, do
    For each training dataset, do
        Propagate the input forward through the network:
        1) Input each $\Delta d_i$ to the network and compute the output $o_u$ of every unit $u$ of the network.
        Propagate the errors backward through the network:
        2) For the network output unit $o$, calculate its error $e_o$
               $e_o = (l - o) * g'(o)$
        3) For each hidden unit $h$, calculate their error values $e_h$
               $e_h = g'(o_h) * \sum_{i \in outputs} (w_{i,h} * e_o)$
           where$g'$ is the first derivative of the sigmoid function.
        4) Update each network weight $w_{j,k}$
               $w_{j,k} = w_{j,k} + \Delta w_{j,k}$
               $\Delta w_{j,k} = r * e_j * x_{j,k}$
```

3.2.3.2 Support Vector Machines for Regression

Support Vector Machine (SVM) is a powerful machine learning algorithm based on minimizing the generalization error bound (structural risk) rather than minimizing the observed training error (empirical risk), so as to achieve better performance. The basic idea of Support Vector Regression (SVR) is to compute a linear regression function in a higher dimensional feature space where the lower dimensional input data are mapped using a kernel function (Basak et al., 2007).

Given training dataset as:

$$\{(x_1, y_1), \dots, (x_\ell, y_\ell)\} \subset \mathcal{X} \times \mathbb{R}$$

where x_i and y_i indicate the attribute and target values respectively, and $\mathcal{X}$ denotes the space of the input patterns (e.g. $\mathcal{X} = \mathbb{R}^d$). In epsilon-SVR, the goal is to find a function *f(x)* that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time as flat as possible. In the simple linear case, *f(x)* has the form as:

$$f(x) = \langle \omega, x \rangle + b \ with \ \omega \in \mathcal{X}, b \in \mathbb{R} \tag{3-7}$$

where $< \cdot, \ >$ denotes the dot product in $\mathcal{X}$, and b indicates a bias value. *Flatness* in Equation (3-7) means seeking a small vector ω. To ensure this, one way is to minimize the Euclidean norm i.e. $\|\omega\|^2 = <\omega, \omega>$. By introducing slack variables ξ_i, ξ_i^* to cope with infeasible constraints in some practical cases or allow for some errors, this problem can be written as the following formulations:

$$\text{minimize} \ \ \frac{1}{2} \parallel \omega \parallel^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*)$$

$$\text{subject to} \ \begin{cases} y_i - \langle \omega, x_i \rangle - b & \leq & \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i & \leq & \varepsilon + \xi_i^* \\ \qquad \xi_i, \xi_i^* & \geq & 0 \end{cases} \tag{3-8}$$

where ξ_i, ξ_i^* denote the allowed upper and lower error bound respectively and the constant C > 0 determines the tradeoff between the flatness of *f* and the amount up to which deviations larger than ξ are tolerated. This corresponds to dealing with the ε-intensive loss function described by Equation (3-9) (Vapnik, 2001):

$$|\xi|_\varepsilon := \begin{cases} 0 & if \ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{3-9}$$

By constructing a Lagrange function and utilizing Lagrange multipliers, the original problem can be solved. The objective function can be rewritten as follows (Vapnik, 2001):

$$f(x) = \sum_{i=1}^{\ell} (a_i - a_i^*) \langle x_i, x \rangle + b \tag{3-10}$$

where α_i, α_i^* are computed Lagrange multipliers. Here, by using a nonlinear kernel function k (x_i, x) satisfying Mercer's condition (Basak et al., 2007) instead of the dot product $<x_i, x>$ in Equation (3-10), SVR can be employed for nonlinear regression.

Support Vector Regression shows two great capabilities that well meet our requirements:

1. SVR is especially suitable for the regression problems for a small sample size. The establishment of facial databases, especially the manual annotation, is an expensive process, therefore it is necessary to maximize the use of limited amount of data.
2. The structural risk minimization principle endows SVR with good generalization capability for unseen data, thus the robustness and adaptation to different subjects of the system are enhanced.

We employ the established LibSVM Library (Chang & Lin, 2011) for the SVR implementation. We apply 16 epsilon-SVRs for the regression of the 16 selected AUs respectively, using the same input/output data format as discussed above. A scaling procedure is also performed before applying SVRs to achieve the best performance. Moreover, kernel selection also plays a key role for the SVR model, since using different kernels may significantly influence the performance when dealing with the same problem. In this research, we consider the non-linear radial basis function (RBF) kernel as a reasonable choice, because:

1. RBF nonlinearly maps inputs into a higher dimensional space, thus it can well handle the case that the relation between facial features and AU intensity levels is nonlinear.
2. RBF has fewer number of hyperparameters than other nonlinear kernels (e.g. polynomial kernel), which may reduce the complexity of model selection (Hsu et al., 2010).
3. RBF usually has lower computational complexity, which in turn indicates better real-time computational performance.

Please note that when the dimensions of features are very high (e.g. thousands), the RBF kernel may become not suitable in comparison to a linear kernel (Hsu et al., 2010). However, it is not the case in this application.

Once the RBF kernel is selected, an essential step is to find optimized sets of cost (C), gamma (g) and epsilon (ε) parameters. We perform a “grid search” procedure on those parameters using the cross-validation technique, since it is regarded as one of the most effective methods to prevent over-fitting (Chang & Lin, 2011). In *v*-fold cross-validation, the overall dataset is firstly divided into *v* groups with equal number of samples in each group, then we use *v*-1 groups of the data for training and the remaining group for testing. This process is repeated *v* times so that each group can be tested in turn. Specifically, various combinations of parameter values (i.e. exponentially growing values: C = 2^{-10}, 2^{-9}, ..., 2^{15}; g = 2^{-15}, 2^{-14}, ..., 2^{10}; ε = 2^{-10}, 2^{-9}, ..., 2^{-1}) are conducted and the one with the lowest Mean Squared Error (MSE) under 5-fold cross-validation is selected. The MSE evaluates the prediction results by taking into account the squared error of the predicted value from the ground truth and can be computed as follows (DeGroot & Schervish, 2011):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{\sim})^2 \quad (3\text{-}11)$$

where y_i is the predicted value, and $y_i^{\sim}$ is the ground truth. Moreover, the Pearson correlation coefficient is also employed to evaluate the linear relationship between the prediction and the ground truth, i.e. how they change together.

Thus, 16 Neural Network and 16 SVR based predictors are implemented to estimate the intensity for each AU respectively. Both manually and automatic selected features are compared and employed as inputs respectively to NNs and SVRs to measure the intensities of 16 AUs. A total of 729 3D facial scans extracted from the Bosphorus database (Savran et al., 2008) from 56 subjects is used for performance evaluation. The databases, experiments and evaluations are detailed in Section 3.4.

3.3 Facial expression recognition using the derived AU intensities

The mapping between AU intensities and emotions could be a challenging task. For example, a ‘surprised’ facial expression may indicate the presence of {AU1, AU2, AU5, AU26}, or the physical cues of {AU1, AU2, AU26} in different cases. The intensities of these present AUs could also be variable. These practical issues make deterministic rule-based techniques less effective (e.g. using translating formula: surprise = AU1+AU2+ AU5+AU26 (Ekman et al., 2002)). Likewise, directly applying machine learning algorithms could be still very challenging, since extensive training data are needed to accommodate various possible combinations of AUs for emotional expressions. There are, however, more than thousands of possible AU combinations in spontaneous facial expressions (Ekman & Friesen, 1983), which are far beyond the data available in any existing databases. In order to deal with such challenges, we propose a novel method to robustly map AU intensities to the six basic emotions using a limited number of samples, which consists of two steps: (1) AU-Emotion relationship mining and ranking; (2) facial expression recognition using the identified discriminative AU combinations.

3.3.1 Mining and ranking AU-emotion relationships

As mentioned above, instead of using the full set of 16 AUs for emotion interpretation indiscriminately, we first derive AU-Emotion relationship, and then identify the AU combinations with the best recognition accuracies as the discriminative AU combinations for each emotion category for subsequent recognition. The AU-Emotion relationship is derived through statistical analysis of sufficient amount of valid samples with AU intensity and emotion annotations provided by the extended Cohn Kanade (CK+) (Lucey et al., 2010) and Bosphorus databases (Savran et al., 2008).

Velusamy et al. (2011) suggested a concept called discriminative power, which applied the probability of an AU, given that a specific emotion has occurred to describe the AU-Emotion relationship. In this research, the AU intensities are described by continuous values rather than only “true” or “false”. Thus, a new concept, Influence Power, is proposed to describe the weights of the AU-Emotion relationship, as defined in Equation (3-12):

$$P = \left(\sum_{i=1}^{n} Intensity_{x,i}\right)/n \qquad (3\text{-}12)$$

where n is the number of examples belonging to a given emotion category, $Intensity_x$ donating the intensity value of AUx occurred corresponding to the given emotion, and the magnitude of P quantifies the Influence Power of AUx for that emotion category. A higher Influence Power represents a closer connection between an AU and an emotion, while a lower value may indicate the weak association between them. 1200 samples (equally distributed to the six basic emotions) collected from the CK+ (Lucey et al., 2010) and Bosphorus databases (Savran et al., 2008) have been taken into account for AU-Emotion relationship identification. After normalizing P across all of the 16 AUs for each emotion, we draw the relation confusion matrix between the 16 AUs and the six basic emotions in Figure 3-8. Thus, a set of association weights between AUs and emotions is established.

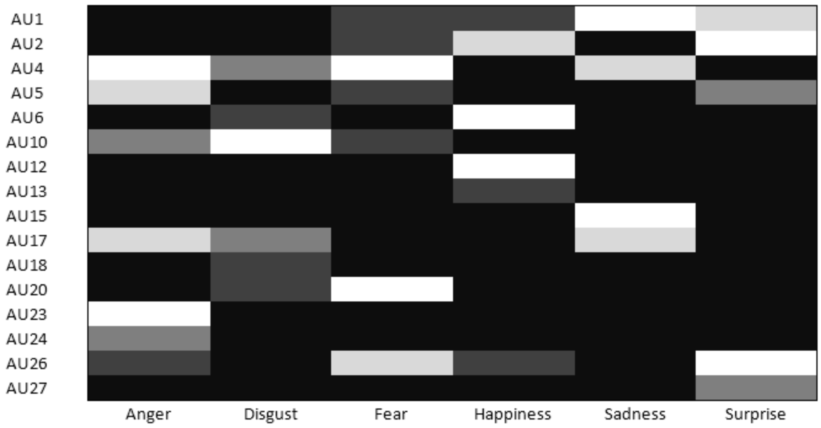


Figure 3-8 The AU-Emotion relation confusion matrix (lighter color indicating higher Influence Power with darker color representing lower Influence Power)

Having obtained the relation confusion matrix, we then select the top ***N*** AUs with the highest Influence Power for the recognition of each emotion. On the positive aspect, this may significantly reduce the potential negative impact of those non-dominant or haphazard AUs and improve classification accuracy. For example, ‘happy’ expressions have AU6, AU12 as highly weighted associations with AU2 as a comparatively lower weighted association, while AU2 is also served as a key physical cue and thus has a higher association weight for ‘surprise’ expressions. However, on the negative aspect, over-filtering those AUs with lower Influence Power may also increase the risk of information loss. Thus, in order to optimize the selection of the ***N*** number of AUs, we perform a series of experiments with different ***N*** number of AUs (i.e. using different numbers of top ranking AUs as inputs) for each emotion category. The AU combinations with the best recognition accuracy will be finalized for subsequent emotion classification. The details are discussed in the following.

3.3.2 Selection of discriminative AU combinations

We employ a unique set of discriminative AUs as inputs for the recognition of each emotion category. The selection of the discriminative AU combinations is detailed as follows: We first perform emotion recognition using different numbers of top ranking AUs (i.e. ***N*** = {2, 3, 4, 5, 6}) as inputs, and record the recognition accuracies in each round. Specifically, for each classifier, we collect 120 samples in total, 50 from the CK+ database (Lucey et al., 2010) and 70 from the Bosphorus database (Savran et al., 2008), covering both positive and negative cases (presence/absence of that emotion) with roughly equal quantities. We also apply a 5-fold cross-validation scheme depending on the sample size. The average cross-validation accuracies obtained by SVM classifiers are summarized in Figure 3-9 (the other classifiers yield very similar patterns, thus are omitted in the Figure).

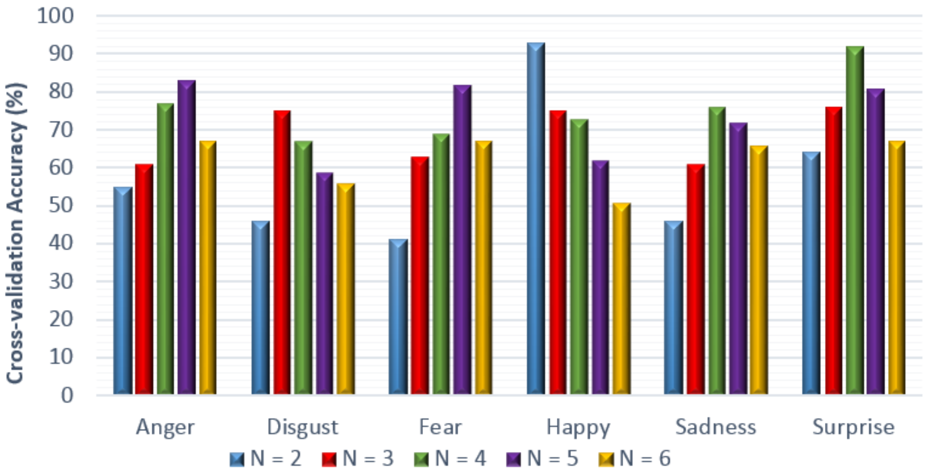


Figure 3-9 Average classification accuracies for the six basic emotions using SVMs and top ranking

AUs (***N*** = {2, 3, 4, 5, 6}) as inputs

Based on the results shown in Figure 3-9, the AU combination leading to the best recognition accuracy is determined as the most discriminative AU combination for each emotion. These AU combinations are summarized in Table 3-3 and employed respectively as the finalized inputs for the six emotion classifiers. For example, in Figure 3-9, since the highest recognition accuracy for ‘anger’ is achieved when ***N*** equals to 5, we select the top five ranking AUs as the discriminative AU combination, i.e. AU4, AU5, AU17, AU23 and AU24. Thus, the derived intensities of these five AUs are subsequently used as inputs to the ‘anger’ emotion classifier. The discriminative AU combinations for other emotion categories are also determined as above. The experimental results and evaluations are presented in Section 3.4.

Table 3-3 Identified discriminative AU sets for the six emotions

Emotions	***Discriminative AU Combinations***				
Anger	AU 4	AU 5	AU 17	AU 23	AU 24
Disgust	AU 4	AU 10	AU 17		
Fear	AU 1	AU 4	AU 10	AU 20	AU 26
Happy	AU 6	AU 12			
Sadness	AU 1	AU 4	AU 15	AU 17	
Surprise	AU 1	AU 2	AU 26	AU 27	

3.3.3 Emotion recognition using adaptive ensemble classifiers

In this research, we propose an adaptive ensemble scheme for the detection of six expressions and any newly arrived novel emotion classes. In this scheme, there are six ensemble classifiers with each ensemble robustly differentiating the presence/absence of each emotion. We also employ single Support Vector Machines (C-SVC) classifiers to conduct the same expression recognition tasks, and their results will be used as the benchmark for comparison with those achieved by the ensemble classifiers.

Ensemble learning generally refers to approaches that generate several base models that are combined to make a prediction, as illustrated in Figure 3-10. Compared to traditional single model-based methods, ensembles have the advantages of improved robustness and increased accuracy (Garcia-Pedrajas et al., 2005). For an exhaustive review of ensemble approaches, readers may refer to Rokach (2010).

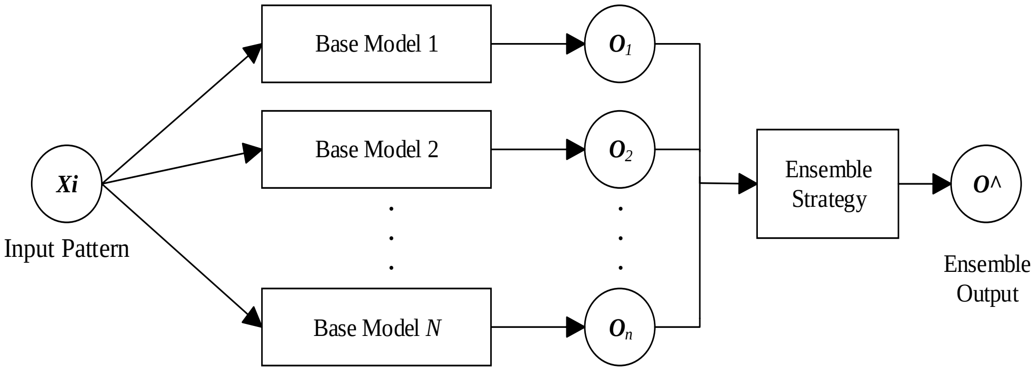


Figure 3-10 An example of an ensemble learning model

In the field of facial emotion recognition, a number of ensemble approaches have been proposed. For example, Whitehill & Omlin (2006) employed the AdaBoost algorithm for AU recognition using Haar features. More recently, Zavaschi et al. (2013) created a pool of base SVM classifiers with features extraction conducted by Gabor filters and Local Binary Patterns, and then applied a multi-objective genetic algorithm to find the best ensemble by minimizing both the error rate and the size of the ensemble. Although ensemble models have been used for facial expression recognition, few of them are capable to detect novel emotion classes.

Moreover, in the field of data stream mining, most of the existing ensemble algorithms integrated with novel class detection employed classic decision tree (e.g. Farid et al., 2013) or k-nearest neighbor (e.g. Masud et al., 2011) classifiers as their base models. In our research, we employ a special type of Neural Network, i.e. Complementary Neural Network, as the base classifier and propose a novel mechanism to further improve the performance of the 6-class emotion recognition and novel emotion detection. The details of our approach are discussed as follows.

Each of the proposed ensemble classification models consists of two phases: **ensemble model generation** (training) and **classification with novel class detection** (testing). Figure 3-11 illustrates the work flow of the generation of an ensemble classifier. It starts with the weight initialization procedure for each training instance based on the posterior probability, as detailed in Section 3.3.3.1. Afterwards, the ensemble model generates a new training subset from the original training set using instances with higher weights. Then, a base model is trained using the newly generated training subset. Here, we employ a novel Complementary Neural Network (CMTNN) as the base classifier, because of its ability to estimate the vagueness level of classification results (see Section 3.3.3.2). The CMTNN is introduced in Section 3.3.3.2. A weight is subsequently calculated and assigned to the current base CMTNN classifier based on its classification accuracy rate for the original training dataset. We also update the weights of the original training instances with the goal of increasing the weights of those misclassified instances. The weight calculation and update methods are discussed in Section 3.3.3.3. The generated training subset is also clustered based on the similarities and differences of the instances, as discussed in Section 3.3.3.4. We employ the following idea for novel emotion class detection. A distance-based clustering technique and the vagueness measure of the classification results obtained by CMTNN will be employed to identify the arrival of novel emotion class (i.e. unseen expressions absent from the training set). Overall, the above procedures iterate three times, thus three weighted base models are generated (considering a balance between performance and computational complexity). The final ensemble classification results can be obtained by using majority of weighted votes of the three base models.

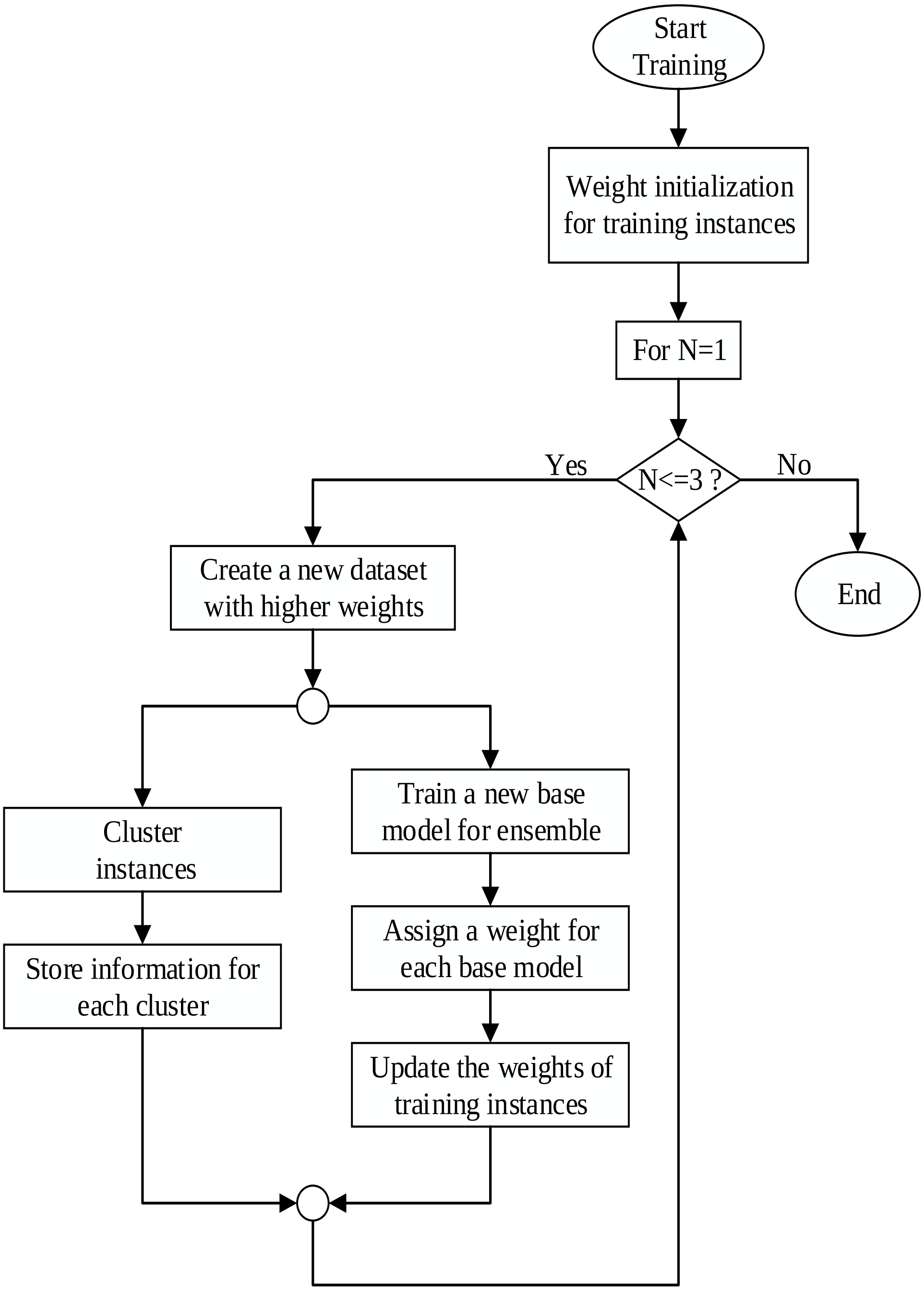

Start Training
Weight initialization for training instances
For N=1
N<=3 ?
Yes
No
End
Create a new dataset with higher weights
Cluster instances
Train a new base model for ensemble
Store information for each cluster
Assign a weight for each base model
Update the weights of training instances

Figure 3-11 Flow chart of the generation of the proposed ensemble model

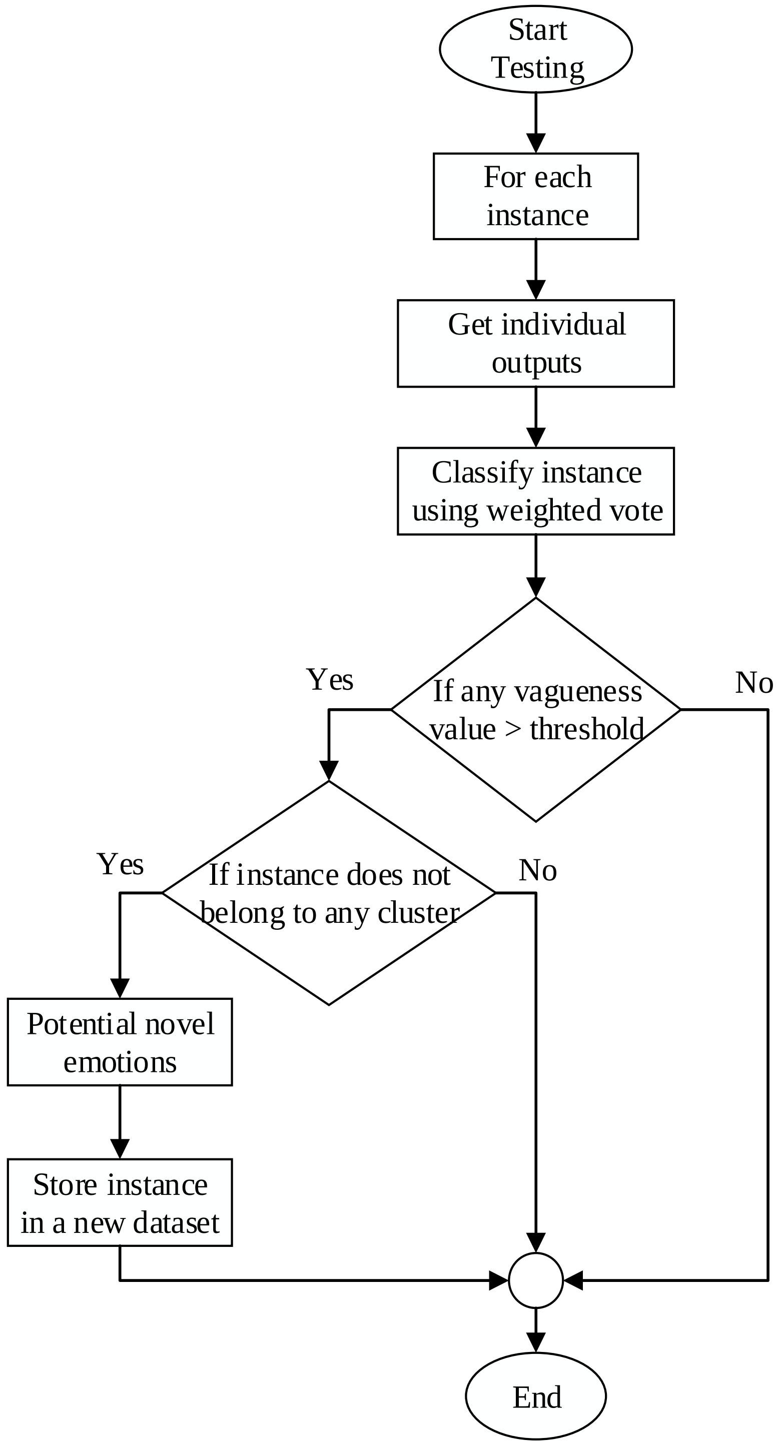

Start
Testing
For each instance
Get individual outputs
Classify instance using weighted vote
If any vagueness value > threshold
Yes
No
If instance does not belong to any cluster
Yes
No
Potential novel emotions
Store instance in a new dataset
End

Figure 3-12 Flow chart of classification with novel emotion detection

Moreover, Figure 3-12 shows the flow chart of classification and novel emotion class detection. As mentioned above, the proposed ensemble scheme is expected to effectively detect novel emotional expressions. Such capability is achieved by the analysis of both the vagueness values of the based models and the corresponding similarity-based clustering results. More specifically, once a testing instance arrives, the three base models for each ensemble respectively output both the individual classification results and the vagueness/uncertainty estimation values of the results (detailed in Section 3.3.3.2). If any of the three vagueness values is greater than a threshold and the instance does not belong to any existing data clusters, then the instance is identified as a potential novel emotion class and will be stored in a separate dataset. Finally, if this instance is identified as a potential novel emotion by more than half of the ensemble classifiers of the six basic emotions (e.g. more than three ensembles), then it is determined as a newly arrived novel emotion.

3.3.3.1 Weight initialization for training instances

First of all, we present the method on how to initialize the weight of each training instance based on naïve Bayes (NB) classifier. Although traditional ensemble approaches (e.g. boosting algorithms) normally initialize the weight of each training instance with an equal value, assigning appropriate weights using non-equal values has been also proved to improve the performance of ensemble classifiers (e.g. Farid et al., 2013).

In this research, the weight of each training instance is initialized based on the posterior probability obtained by a NB classifier. Specifically, we first estimate the prior probability $P(C_i)$ for each class C_i, by calculating how often each class occurs in the given training dataset. Similarly, for each attribute A_j and each class C_i, the class conditional probability $P(A_j|C_i)$ can be obtained by counting how often each attribute value occurs in each class. Given an instance x_i, assuming all attributes are independent, the conditional probability $P(x_i|C_i)$ can be estimated by combining the effects of each different attribute as shown in Equation (3-13):

$$P(x_i|C_i) = \prod_{j=1}^{n} P\left(A_j|C_i\right) \quad (3\text{-}13)$$

Then, the posterior probability $P(C_i|x_i)$ can be calculated according to Bayes' theorem as:

$$P(C_i|x_i) = \frac{P(x_i|C_i)P(C_i)}{P(x_i)} \tag{3-14}$$

Thus, the posterior probability is obtained for each class. We then assign a weight for the instance x_i using the highest posterior probability. The weights of the rest instances are initialized using the same method. Once the weights of all instances are initialized, their weights will be normalized so that their sum equals to 1.

3.3.3.2 Base model generation (CMTNN)

Having initialized the weight for each training instance, we focus on the generation of each base model. Here, we introduce a CMTNN as the base classifier, which is not only especially suitable for binary classification problems, but also able to provide vagueness estimation of the classification results.

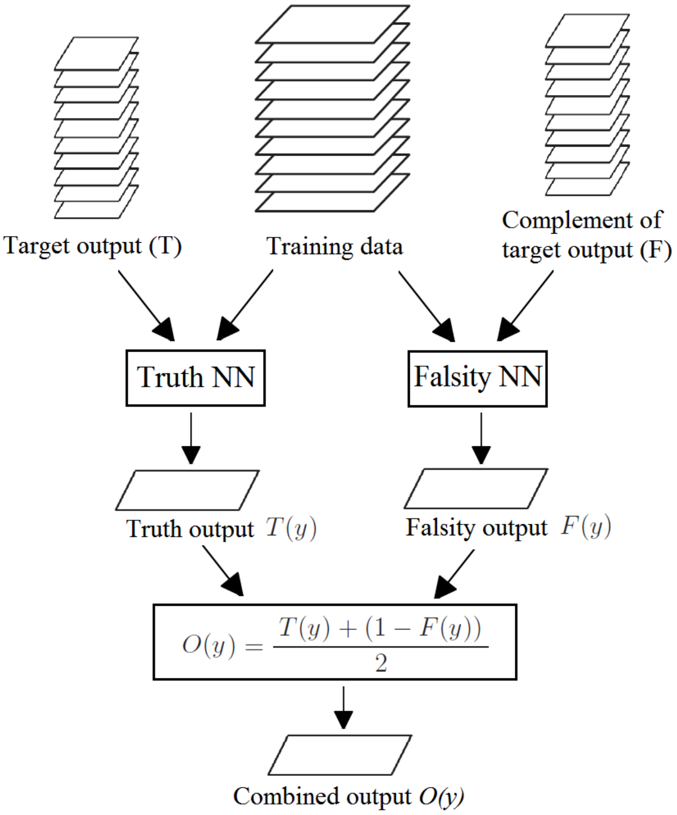


Figure 3-13 Topology of a Complementary Neural Network (Kraipeerapun, 2008)

CMTNN, originally proposed by Kraipeerapun (2008), consists of a pair of opposite feedforward Neural Networks with the same architecture (i.e. a truth Neural Network and a falsity Neural Network). The truth Neural Network is trained by original training data to predict the degree of the truth membership values, and the falsity Neural Network is trained to predict the degree of the false membership values using the same inputs but the complement of target outputs of the original training instances (as illustrated in Figure 3-13). For instance, if the target output of original training data is 1, the complement of this target output used to train the falsity Neural Network should be 0.

For each test pattern, a CMTNN outputs both the truth and false membership values, and they are supposed to be complementary to each other ideally (i.e. if the truth membership value is 1 then the false one is supposed to be 0, or vice-versa). In practice, however, both membership values predicted may not always be informative enough for the final classification. For example, both the truth and false membership values are around 0.5. Thus, an uncertain classification occurs. Empirically, the greater proximity of the truth and false membership values, the higher the degree of vagueness exists. Given a testing pattern, let y_i be the output. $T(y_i)$ denotes the truth membership output, and $F(y_i)$ denotes the false membership output, then the vagueness value of the prediction $V(y_i)$ can be estimated as:

$$V(y_i) = 1 - |T(y_i) - F(y_i)| \tag{3-15}$$

By combining $T(y_i)$ and the complement of $F(y_i)$ using a simple equal weighted method, the final output $O(y_i)$ for the pattern can by calculated as:

$$O(y_i) = \frac{T(y_i)+(1-F(y_i))}{2} \tag{3-16}$$

A threshold value is applied to Equation (3-16) to classify the output into binary classes (generally, the most commonly used threshold value is 0.5). An output pattern is classified as 1 (true) if $O(y_i)$ is greater than the threshold value, otherwise, it is classified as 0 (false). Compared to other traditional methods which solely apply truth membership values, CMTNN has two outstanding features: improved classification accuracy for binary problems and the ability to assess uncertainty of classification using the vagueness value (Jeatrakul & Wong, 2009).

3.3.3.3 Weight calculation and update

We then introduce the weight calculation methods for both of the base classifiers and training instances. First, once a base classifier is generated, a weight will be assigned based on its classification accuracy rate for the original training instances. Once all the three classifiers are generated, their weights will be normalized so that their sum equals to 1.

Moreover, for training instances, we apply the following steps to update their weights, with the intention to increase the weights of those instances which are more difficult to classify (i.e. those with higher error rates). We first assign an error rate for each training instances x_i by:

$$error(x_i) = \begin{cases} 1, & if\ misclassified \\ 0, & if\ correctly\ clasified \end{cases} \tag{3-17}$$

We then calculate the overall error rate for all instances as follows:

$$error_{overall} = \sum_{i=1}^{n} w_i * error(x_i) \tag{3-18}$$

where w_i is the current weight for instance x_i. Afterwards, the weights of the correctly classified instances will be decreased by Equation (3-19):

$$w_{i,updated} = w_i * (\frac{error_{overall}}{1-error_{overall}}) \tag{3-19}$$

Thus, the weights of correctly classified instances are decreased and the weights of those misclassified ones become increased comparatively. Once the weights of all instances are updated, their weights will be normalized, so that their sum remains the same as it was before.

3.3.3.4 Distance-based data clustering

Clustering is a widely-used unsupervised learning technique. It is a main task of exploratory data mining, and has been applied to many application domains such as image analysis, pattern recognition, information retrieval, medicine, and bioinformatics. It is a form of learning by observation, and aims to determine the intrinsic grouping for a set of unlabeled data based on the principle that instances in the same group (called a cluster) are similar (or related) to each other and different from (or unrelated to) the instances in other groups. The greater the difference between clusters, and the greater the similarity within a cluster, the better the clustering.

In the distance-based clustering, we use the Euclidean distance as the metric to determining the similarity (or differences) of two instances. For a given instance x_i, if we can find any instance x_j in an existing cluster N that fulfills: (1) the Euclidean distance $D_{i,j}$ between x_i and x_j is minimum, and (2) $D_{i,j}$ < a predetermined threshold, the instance x_i is assigned to N. Otherwise, x_i is assigned to a newly generated or any other cluster. During the training phase, the distance-based clustering is employed to specially measure the distribution of the training instances. During the testing phase, if the output uncertainty level (i.e. the vagueness value of a CMTNN) of an instance is greater than a predetermined threshold, this instance will be further determined by the distance-based clustering. If the instance does not belong to any existing clusters, it is confirmed as a potential novel class.

3.4 Evaluation and discussion

In this section, we perform two types of evaluations of the proposed system: static off-line and real-time on-line evaluations. The off-line evaluation is purely based on annotated facial images borrowed from the Bosphorus database, for which we conduct exhaustive experiments for both AU intensity estimation and emotion classification to evaluate the system performance. The on-line testing mainly focuses on the assessment of the system's real-time performance and newly arrived novel emotion class detection, where we use the system trained with the database images to recognize facial expressions of real human subjects in real time.

3.4.1 Facial expression databases

In this research, we employ two facial expression image databases. The first database employed is the CK+ database, which is based on 2D facial images but provides rich AU intensity and expression annotations. However, this database is only used for the statistical computation of the discriminative AU sets for each emotion as discussed in Section 3.3. The second database employed for this research is the Bosphorus 3D Database, which contains both 3D facial scans and manually labeled landmarks, as well as a large variety of Action Unit and expression annotation. This database is used for the evaluation of both AU regression and emotion classification. The introduction of these two databases is provided in the following:

- **The Extended Cohn-Kanade Database** consists of 593 image sequences across 123 subjects with each image sequence starting from a neutral expression and ending in a peak frame emotional expression. Among 593 image sequences, the annotations of the six basic emotions and facial AUs are provided for 327 peak frame images. The AU annotations in the CK+ database have been provided with a numbered scale from 1 to 5 and hence the target intensity values in the range levels of A – E are accordingly scaled. These AU intensity and expression data are used only for the AU-Emotion Relationship analysis and discriminative AU set selection, as detailed in Section 3.3.1.
- **The Bosphorus 3D Database** includes a rich set of 4652 3D facial scans and corresponding manually labeled facial landmarks collected from 105 subjects (including 60 men and 45 women; 29 of them are professional actors/actresses). Both Action Units (25 out of the 44 defined in FACS) and the six basic emotions are annotated specifically for the purposes of facial expression analysis. The 3D facial scans are acquired by Inspeck Mega Capturor II 3D, with about 0.3mm depth resolution in *x*, *y*, and *z* dimensions and 1600x1200 pixels high color texture resolution (Savran et al., 2008). In this study, excluding occlusion facial scans, a

subset of the database containing clear annotation for both AU intensity and the six basic emotions is selected. The subset includes 729 facial scans covering 56 subjects, and we extract a total of 960 samples for the evaluation of the intensity estimation of the 16 AUs (a scan can contain more than one AUs). These scans contain both frontal and non-frontal head poses with yaw rotations from 0 to 30 degrees and pitch rotations ranging from slight upwards, neutral, to slight downwards.

3.4.2 Off-line evaluation

In off-line evaluation, we assess the system's performance by using database sample images with AU intensity and emotion annotations. All the results are obtained using the cross-validation technique. The setting of the off-line evaluation is described in the following:

- For the off-line evaluation, both the training and testing phases were purely based on database images. Therefore, we did not use the Kinect for this evaluation.
- We apply *n*-fold cross-validation to evaluate the performance of both AU intensity estimation and emotion classification, which embeds training and testing phases of the system together. As detailed before, the cross-validation process uses *n* -1 groups of the data for training and the remaining group for testing. This process is repeated *n* times. There are overall 729 FACS coded emotional facial images across 56 subjects borrowed from the Bosphorus 3D Database employed for the cross-validation evaluation for both AU intensity estimation and emotion classification. Specifically, we employ 5-fold cross-validation in our work according to the sample size.
- The computational cost of the learning stage in each round of the cross-validation process is approximately 2-5 seconds for AU intensity estimators on average, and 4-6 seconds for emotion classifiers (such as the ensemble classifiers) on average. The computational cost of the test stage in each round of cross-validation process

is approximately 100-200 milliseconds.

3.4.2.1 Evaluation on AU intensity estimation

As mentioned before, a total of 729 FACS coded emotional facial scans across 56 subjects extracted from the Bosphorus 3D Database (Savran et al., 2008) is used for the evaluation of AU intensity estimation and subsequent emotion classification. The features we used for AU intensity estimation are solely based on the differences of the extracted Euclidean distance features between the neutral and any expressive frames. They are either generated by the manual selection or the mRMR based optimization. For each AU, we have collected around 60 samples, covering both positive cases, i.e. AU presence at any intensity levels (approximately 75%) and negative cases, i.e. AU absence (approximately 25%). A single output value ranging from 0 to 1 is used to represent AU absence through maximum intensity. We apply the 5-fold cross-validation as described above to evaluate the prediction accuracy and generalization capability for each AU. The output AU intensities are subsequently compared against the ground truth to calculate the MSE and CORR for each AU.

In the existing research of AU recognition, the accuracy tends to heavily depend on the training sample size. Typically, most of them required a large number of training images (e.g. thousands) with good diversity and coverage to maintain sufficient accuracy and robustness (e.g. Koelstra et al., 2010; Whitehill et al., 2011; Savran et al., 2012). In order to deal with such challenges, we employ the most discriminative motion-based facial features which enable a significant reduction of training data for AU intensity estimation and in the meantime provide an impressive performance. As shown in Figure 3-14, the average MSE for SVR based AU intensity estimation remains stably below 0.1 once the sample size reaches approximately 50.

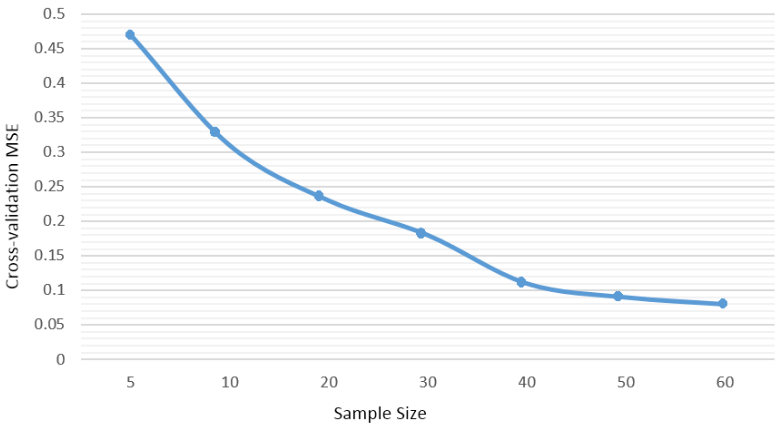


Figure 3-14 Average cross-validation MSE for AU regression in relation to the data sample size used

Using manually selected features

First of all, Table 3-4 shows the results obtained by the feedforward Neural Networks (BPNNs) and Support Vector Regressors (SVRs) for AU intensity estimation using manually selected features. For both BPNNs and SVRs, the lowest MSEs (below 0.05) are observed for AU13 (Cheek Puffer), AU2 (Outer Brow Raiser), AU26 (Jaw Drop), AU10 (Upper Lip Raiser), AU12 (Lip Corner Puller) and AU17 (Chin Raiser) followed by AU1 (Inner Brow Raiser), AU15 (Lip Corner Depressor), AU20 (Lip Stretcher), AU18 (Lip Puckerer), AU4 (Brow Lowerer), AU23 (Lip Tightner) and AU27 (Mouth Stretch), which also obtain fairly low MSEs below 0.1. These results demonstrate the effectiveness and robustness of the extracted motion-based facial features for AU intensity regression.

In contrast, relative higher MSEs (above 0.1) are also observed for the intensity estimation of some AUs, such as AU6 (Cheek Raiser), AU5 (Upper Lid Raiser) and AU24 (Lip Pressor). These results can be explained by the fact that the facial movements of these AUs are very subtle. Especially for AU24, which has the highest MSE and lowest CORR. It could be attributed to the reason that both AU23 and AU24 can cause similar lip boundary changes (e.g. the red parts of lips are narrowed), which may lead to ambiguous annotations even for expert coders. On average, BPNNs and SVRs yield similar performances for AU intensity estimation. However, SVRs are found to perform better than BPNNs for more subtle AUs, in terms of both MSE and CORR measurements (e.g. AU5, AU6 and AU24).

Table 3-4 Results for AU intensity estimation using manually selected features (BPNN= Backpropagation Neural Network, SVR=Support Vector Regression)

AUs	*MSE (%)*		*CORR*	
	BPNN	***SVR***	***BPNN***	***SVR***
AU 13	1.1	2.0	0.952	0.957
AU 2	1.3	2.7	0.970	0.978
AU 26	2.5	3.1	0.954	0.976
AU 10	3.6	3.3	0.924	0.939
AU 12	4.2	3.9	0.939	0.930
AU 17	4.3	4.1	0.896	0.923
AU 1	4.7	5.1	0.957	0.960
AU 15	5.6	6.0	0.890	0.892
AU 20	5.8	4.6	0.878	0.913
AU 18	6.4	5.6	0.955	0.947
AU 4	6.6	5.9	0.893	0.824
AU 23	9.2	9.9	0.921	0.925
AU 27	9.7	10.4	0.931	0.969
AU 6	11.9	10.7	0.841	0.859
AU 5	13.4	12.3	0.881	0.895
AU 24	14.9	12.6	0.790	0.863
Overall	**6.5%**	**6.3%**	**0.911**	**0.921**

Using automatically selected features

Next, we employ the automatically selected features obtained by using the mRMR-based optimization to estimate the intensities of the 16 selected AUs. The results obtained are summarized in Table 3-5. Empirically, a few informative features with great discrimination power (i.e. 10 to 20 features in general) are sufficient to yield good results. On average, the automatically selected features achieve comparable performance in comparison to the manually selected features for the intensity estimation for many AUs (e.g. AU2, AU13, AU15, AU26, and AU27). For some AUs, such as AU2 and AU13, the automatic features generate even lower MSE values when SVRs are used. However, for some other AUs, such as AU4, AU20 and AU24, the performance drops slightly in comparison to the manual feature selection. Overall, the mRMR-based feature selection yields a very close performance to the manually devised features in terms of both averaged MSE and CORR values. Thus, the AU intensities obtained by SVRs with the corresponding automatically selected features as inputs will be used for subsequent emotion recognition.

Furthermore, since all the results are obtained in the form of continuous AU intensity levels, they reflect more physical truth of facial expressions in comparison to other applications that only performed presence or absence binary-classifications (e.g. Tsalakanidou & Malassiotis, 2010; Li et al., 2013). Such AU intensity measurements may also indicate effective physical cues to contribute to the sequent emotion classification.

Table 3-5 Results for AU intensity estimation using automatically selected features (BPNN= Backpropagation Neural Network, SVR=Support Vector Regression)

AUs	*MSE (%)*		*CORR*	
	BPNN	*SVR*	*BPNN*	*SVR*
AU 2	1.3	1.7	0.937	0.953
AU 13	2.1	1.4	0.919	0.975
AU 26	3.2	3.1	0.923	0.975
AU 10	3.9	4.1	0.885	0.938
AU 12	5.9	5.3	0.895	0.926
AU 17	5.7	5.9	0.873	0.900
AU 1	6.6	6.0	0.906	0.936
AU 15	6.6	6.2	0.874	0.891
AU 20	5.9	6.4	0.875	0.912
AU 18	7.7	6.9	0.911	0.936
AU 4	8.0	7.8	0.897	0.805
AU 23	9.5	9.4	0.893	0.905
AU 27	10.2	9.7	0.886	0.963
AU 6	12.0	11.7	0.822	0.838
AU 5	13.6	13.3	0.831	0.878
AU 24	15.2	14.2	0.787	0.857
Overall	**7.3%**	**7.1%**	**0.882**	**0.912**

3.4.2.2 Evaluation on facial expression recognition

The 729 facial scans used for AU intensity estimation above are then applied for the evaluation of the facial emotion recognition. As mentioned earlier, the intensities of the 16 diagnostic AUs generated by SVRs with mRMR based feature selection are subsequently used as inputs to the six ensemble classifiers for facial expression recognition. Six single SVM classifiers are also used to perform facial expression recognition for the comparison with the ensemble classifiers. We also apply a 5-fold cross-validation to measure the accuracy performance of each emotion recognition classifier. We measure the performance of the proposed emotion recognition approaches in terms of the accuracy confusion matrix and F1-measure. A confusion matrix is a $n * n$ matrix, where the row labels are ground-truth emotion annotations and the column labels are the classification results. The diagonal entries indicate the correct classifications, while the off-diagonal entries correspond to misclassifications. The F1-measure is a harmonic mean of precision and recall rate, which is considered to be a more comprehensive metric.

Table 3-6 presents the recognition accuracy confusion matrices for the six basic emotions obtained by SVMs and the proposed ensemble classifiers. Overall, by using SVMs for emotion classification, we achieve an average recognition accuracy rate of 90.5% (shown in Table 3-6 (a)), while by using ensemble models, we obtain a higher overall accuracy of 92.2% (see Table 3-6 (b)). More specifically, for either approach, the best performances are achieved for the recognition of 'happy' and 'surprised' facial expressions, with recognition accuracies beyond 95%. For 'anger' and 'fear', slightly lower recognition accuracies are observed for both approaches with the ensembles (92.8% for 'anger' and 92.1% for 'fear') outperforming the SVM classifiers (91.3% for 'anger' and 91.1% for 'fear').

Table 3-6 Confusion matrices of facial emotion recognition accuracies

	Anger	***Disgust***	***Fear***	***Happiness***	***Sadness***	***Surprise***
a. Recognition accuracy (average 90.5%) using SVM classifiers						
Anger	91.3	6.7	0	0	13.6	0
Disgust	11.1	85.6	4.3	0	0	0
Fear	0	0	91.1	0	0	11.8
Happy	0	0	0	95.6	0	8.8
Sadness	5.8	3.1	9.8	0	82.7	0
Surprise	0	0	0	7.9	0	96.5
b. Recognition accuracy (average 92.2%) using the proposed ensemble classifiers						
Anger	92.8	3.3	0	0	9.3	0
Disgust	9.8	88.6	2.3	0	0	0
Fear	0	0	92.1	0	0	9.9
Happy	0	0	0	96.1	0	8.9
Sadness	4.7	0	7.3	0	86.6	0
Surprise	0	0	0	7.3	0	96.7

For ‘disgust’, a lower recognition accuracy of 85.6% is observed when using the SVMs, and 88.6% when using the ensembles. A possible explanation is that those emotions with comparatively lower recognition accuracies often entangled with more complicated and subtle facial changes than the ones with higher recognition accuracies, and thus more challenging to recognize. The lowest recognition rates are observed for ‘sadness’ (82.7% by SVM and 86.6% by the ensemble classifier). This could be due to the fact that in some facial scans, subjects inaccurately express ‘sadness’ using the combination of AU20 (Lip Stretcher) and AU15 (Lip Corner Depressor), rather than solely using AU15 as indicated by FACS (Ekman et al., 2002). But AU20 is also served as a key physical cue for ‘fear’, which may lead to misclassification of ‘sadness’ as ‘fear’.

Table 3-7 Comparison of recognition accuracies for the six basic emotions

	Accuracy _SVM	***Accuracy _Ensemble***	Salahshoor & Faez (2012)	Ujir (2013)
Surprise	96.5	96.7	91.4	90.8
Happiness	95.6	96.1	74.3	100.0
Fear	91.1	92.1	92.9	21.5
Anger	91.3	92.8	87.3	75.4
Disgust	85.6	88.6	78.3	43.1
Sadness	82.7	86.6	95.5	67.7
Overall	90.5%	92.2%	86%	66.4%

We subsequently compare our work with other state-of-the-art developments such as Salahshoor & Faez (2012) and Ujir (2013) in Table 3-7. These related applications are chosen because of their focus on a similar research challenge of 3D facial emotion recognition and the employment of the same Bosphorus 3D database and similar evaluation strategies. The comparison in Table 3-7 indicates that our proposed facial emotion recognition system outperforms both of the above related developments. Specifically, the 'surprised' facial expression has been well recognized by all the three systems (accuracies > 90%). However, the two related systems also respectively show considerable limitations for the recognition of the other emotion categories. For example, the system of Salahshoor & Faez (2012) performed poorly for the recognition of 'happy' and 'disgust' (accuracies < 80%) emotions, whereas the work of Ujir (2013) also indicated very unstable classification performance for 'fear' (only 21.5%) and 'disgust' (43.1%) expressions. In comparison to these state-of-the-art applications, our system is proved to be more stable for the recognition of all of the six emotion categories and achieves the highest overall recognition accuracy among the related applications.

Table 3-8 Comparison of F1-measures for the six basic emotions

	F1_SVM	***F1_ Ensemble***	***F1_*** Sandbach et al. (2012)
Surprise	0.889	0.897	0.826
Happiness	0.94	0.945	0.812
Fear	0.888	0.913	0.462
Anger	0.877	0.895	0.500
Disgust	0.876	0.923	0.644
Sadness	0.843	0.884	0.625
Overall	0.89	0.91	0.65

Since the classification accuracy rate could be less informative sometimes, especially when the data is unbalanced, the F1-measure for each emotion category is also presented in Table 3-8. We also compare our system with the work by Sandbach et al. (2012) because of their state-of-the-art performance and the employment of the same performance metric (i.e. the F1-measure). Based on the comparison of the F1-measure results, it is noticed that the performance of our system significantly outperforms those of the work by Sandbach et al. (2012). Although their HMM based approach also generated good results for the recognition of 'happy' and 'surprised' facial expressions, our system performs more stably for the detection of each emotion category.

Overall, the above results demonstrate that the proposed system is consistently an efficient and robust solution for AU intensity estimation and emotion recognition. Furthermore, facial expressions sometimes may contain a mixture of emotions, thus it is possible that two (or more) emotional states occur simultaneously in one emotional facial scan. The proposed approach also shows great potential to detect such combination of emotions (e.g. happy + surprise) by deriving recognition results for each emotion category separately.

3.4.3 On-line evaluation

The facial emotion recognition system has also been applied to real-time emotion detection tasks contributed by test human subjects. The facial feature point localization of our system is able to integrate both color and 3D depth image data so that it provides great robustness against illumination changes and pose variations. It thus lays solid foundations for subsequent AU intensity measurement and emotion recognition. Moreover, the computational complexity of the face tracking and landmark localization requires 20-30 milliseconds under normal lab lighting conditions. The mRMR-based feature selection, AU regression, and emotion classification take an averaged run time of 3-5 milliseconds (which may change slightly depending on different types of regressors and classifiers used). Overall, the system is able to perform efficiently for facial emotion recognition at a frame rate of 25~30 fps on i7 4700MQ quad-core CPUs with 8GB RAM.

For the on-line evaluation, our system has been trained with database images first and then is used to recognize human subjects' facial expressions in real time. The setting of the online testing is provided in the following:

- In the online evaluation, our system has been trained with database images first. Then the Kinect is used in the testing phase to track human subjects' facial landmarks. Based on the tracked facial landmarks, the system subsequently performs feature extraction and selection, AU intensity estimation and emotion recognition.
- In the on-line evaluation, the above 729 FACS coded database images from 56 subjects employed for the off-line evaluation are entirely used for training of both the AU intensity estimators and emotion classifiers. The training computational cost of the system is approximately 4-5 seconds for AU intensity estimators while 5-7 seconds for emotion classifiers.
- For on-line testing, we recruit eleven participants with five females and six males aging from 25 to 40 years old. Majority of them are postgraduate students and all the test subjects are non-experts in the field. The computational cost of the system in the real-time testing is about 3-5 milliseconds per frame.

As mentioned above, we recruit eleven participants for real-time system evaluation. In order to ensure effective tracking of facial geometric features, the distance between the participants and the Kinect was controlled within the range of 2 (±0.5) meters. The participants were required to display a series of emotional clips. Each clip lasts approximately 10–15 seconds (i.e. 300–450 frames). It starts from a short neutral state period (4–5 seconds) and followed by a posed facial expression period. Both the neutral state and expression periods were manually labeled in each clip by an expert annotator. In addition to the six basic emotions (happiness, sadness, disgust, surprise, fear and anger) that are collected from the test subjects and used to test the system, we also evaluate the system with some novel emotional expressions (e.g. contempt and excitement) contributed by the test subjects.

In our experiment, the expressions of 'contempt' emotion require a subject to show the facial behavior of dimpler (AU14) while the expressions of 'excitement' emotion require the combination of 'surprise' and 'happy' expressions with the upper face showing inner and outer brow raiser and upper lid raiser and the lower face indicating cheek raiser and lip corner puller. We use the above guidance for the posing and collection of these two novel emotion classes for testing. Figure 3-15 shows examples of the six basic emotions plus 'contempt' and 'excitement' expressions posed by two test subjects during testing. Eventually, the system was evaluated with a total of 136 emotional clips. The detailed results and discussions are presented as follows.

Figure 3-16 shows an example of real-time detection of a 'surprise' emotional clip using the six ensemble classifiers. The vertical axis indicates the emotion detection results from absence (0) to maximum presence (1) of the 'surprise' expression, and the horizontal axis marks the timeline (in frames). As illustrated in Figure 3-16, for the recognition of 'surprise', ideally, only the corresponding ensemble classifier for 'surprise' generates an output curve consistent with the ground truth. The outputs of the other five ensemble classifiers consistently remain in a much lower level. Overall, the average classification accuracy rate for this emotion clip is 93.2%.

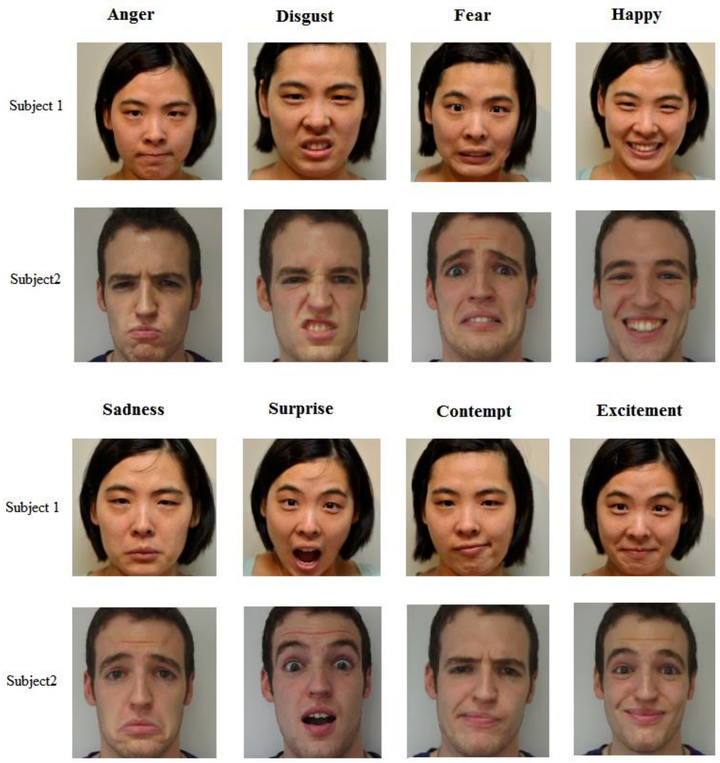


Figure 3-15 Snapshots of the six basic emotions plus ‘contempt’ and ‘excitement’ posed by two test subjects in the on-line evaluation

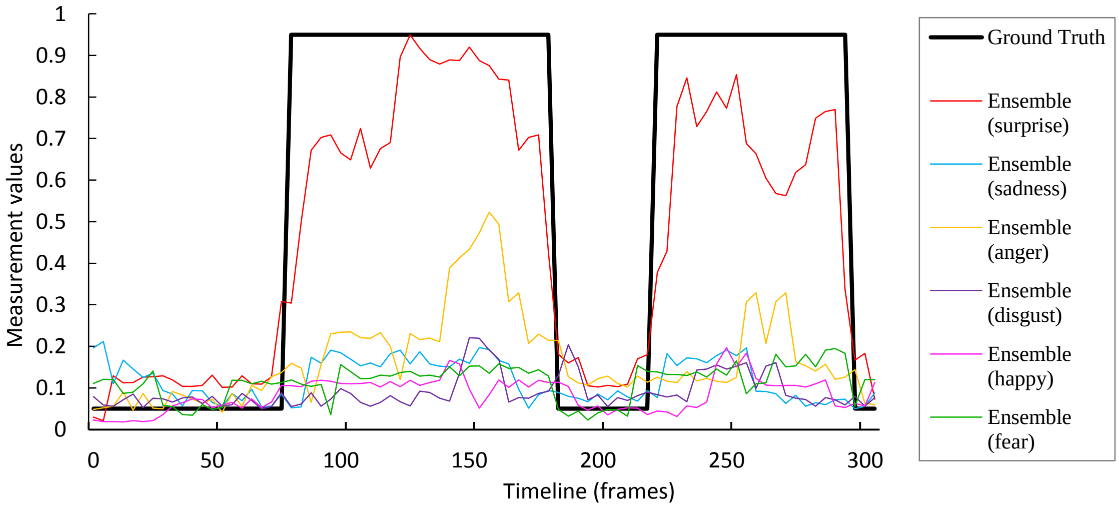


Figure 3-16 Examples of real-time detection of ‘surprise’. The bold black line indicates the ground-truth (presence/absence), and the six color lines respectively indicate the real-time outputs of the six ensemble classifiers

Table 3-9 Real-time recognition accuracies for the six basic emotions and novel emotion classes

	Recognition Accuracy (average 84%)
Surprise	93.2
Happiness	88.1
Fear	81.6
Anger	79.4
Disgust	83.7
Sadness	77.9
	Classified as a Novel Emotion (average 72.2%)
Contempt	77.2
Excitement	67.1

Table 3-9 summarizes the real-time recognition accuracy rates for the six basic emotions and novel emotion detection rates for ‘contempt’ and ‘excitement’. Generally, the on-line system yields comparable results to those obtained in off-line evaluation. Except for ‘anger’ and ‘sadness’, the recognition accuracy rates for the other four basic emotions are consistently beyond 80%, which only show a slight decrease compared to the results obtained in previous off-line evaluation. More important, 77.2% of ‘contempt’ and 67.1% of ‘excitement’ expressions are successfully identified as novel emotion classes rather than only roughly classified as one of the six basic emotions. These results demonstrate that the proposed ensemble classifiers are well capable of detecting newly arrived novel emotion categories and show great improvements compared to other existing systems.

3.5 Summary

In this chapter, we presented a fully automatic system for real-time 3D AU intensity estimation and emotion recognition. We first realized real-time 3D face tracking and facial landmark extraction based on the Kinect platform. Then 16 sets of motion-based facial features containing rich person-independent emotional information were extracted and selected by using both manual and mRMR-based automatic feature selection methods. These feature sets were subsequently employed as inputs to an array of Neural Networks and Support Vector Regressors respectively to estimate the intensities of the 16 diagnostic AUs. Experimental results indicated that the mRMR based optimized feature selection yields comparable results in comparison to the manually selected features when using either Neural Networks or SVRs for AU intensity measurement. Moreover, the SVR-based AU intensity estimation slightly outperformed the Neural Network based method. This is probably caused by the fact that the grid search with cross validation has been conducted for optimal parameter selection for the SVR models. By using the automatically selected features and SVRs, we have achieved an averaged MSE of 0.071 and an averaged CORR of 0.912 for the intensity estimation of the 16 AUs. The intensities of AU2 (Outer Brow Raiser), AU10 (Upper Lip Raiser), AU13 (Cheek Puffer) and AU26 (Jaw Drop) were well estimated with lowest errors (MSE < 0.05), whereas more subtle AUs, such as AU5 (Upper Lid Raiser), AU6 (Cheek Raiser), and AU24 (Lip Pressor) were estimated with relatively higher estimation errors (MSE > 0.1). The above results also demonstrated the extracted motion-based facial features are very efficient and robust for AU intensity estimation.

We subsequently used the derived AU intensities to recognize the six basic emotions using the identified discriminative AU combinations and dedicated ensemble classifiers for each emotion category. The proposed novel adaptive ensemble classifiers show great robustness and flexibility for not only the recognition of six basic emotions but also the detection of newly arrived unseen novel emotion categories. The off-line evaluation results using the Bosphorus database indicated that the proposed ensemble models consistently outperform the SVM-based classification, and have achieved an averaged recognition accuracy of 92.2% and an averaged F1-measure of 91% for the recognition of the six basic emotions. The best recognition accuracies were obtained for ‘happy’ and ‘surprise’ facial expressions (> 96%) with ‘fear’, ‘anger’ and ‘disgust’ reasonably recognized (>88%). The lowest recognition accuracy rate was observed for ‘sadness’ (86.6%). The system also outperforms other state-of-the-art research on 3D facial emotion recognition tasks based on the comparison of both the recognition accuracy and F1-measure results.

We also conducted an on-line evaluation with real human subjects to assess the system’s real-time performance and the efficiency for novel emotion class detection. Overall, the proposed system is able to perform facial emotion recognition efficiently with a frame rate of 25~30 fps on i7 4700MQ quad-core CPUs with 8GB RAM. We obtained an impressive average recognition accuracy rate of 84% for the detection of the six expressions when tested with real human subjects (only slightly lower than those achieved in off-line evaluation). Moreover, the proposed ensemble classifiers also show superior ability to detect the arrival of novel emotion classes with 72.2% detection rate on average.

Chapter 4 Dimensional emotion regression for whole-body expressions

In this chapter, we address the problem of real-time continuous regression of users' emotional states in a valence and arousal space based on their whole-body expressions. That is the proposed system maps subjects' emotional states to a two-dimensional coordinate space spanned by arousal and valence, where each value ranges between -1 and 1. First of all, we systematically consider and extract users' static and dynamic bodily features. Genetic Algorithm (GA) optimization is then employed to conduct feature selection and identify their most optimal discriminative combinations for affective dimensional regression. We also examine how both static and dynamic features perform for the regression of each affective dimension. In order to robustly predict users' continuous affective dimensions in the valence and arousal space, we propose a novel ensemble regression model with great adaptability to deal with newly arrived unseen bodily expressions and data stream regression. Additionally, as pointed out by Kleinsmith & Bianchi-Berthouze (2013) and Metallinou et al. (2013), continuous and dimensional affective annotation is inherently a challenging task. We present a novel annotation method based on inter-annotator correlations and mean value differences to effectively fuse multiple annotations to build ground truth for system evaluation.

The remainder of this chapter is organized as follows: Section 4.1 presents feature extraction from whole-body expressions and automatic feature selection using the GA. The proposed adaptive ensemble model for continuous and dimensional affect regression is subsequently discussed in Section 4.2, together with the other two benchmark single regression methods. In Section 0, we discuss the process of data collection and affective annotation, as well as experimental results in comparison with other state-of-the-art research. Finally, we draw conclusions in Section 4.4.

4.1 Feature extraction and selection

In this section, we first of all discuss the method about how the body expression information is captured and extracted. We then present the GA-based automatic feature selection in order to identify the most optimal and discriminative combination of static and dynamic features for the interpretation of each affective dimension.

4.1.1 Whole-body expression feature extraction

In this research, we use Microsoft Kinect and its Natural User Interface SDK (Microsoft Corporation, 2013) to recognize users and track their bodily behaviors in real-time. The Kinect provides an effective and economical way for 3D body information tracking, as mentioned earlier which physically contains an RGB camera, a multi-array microphone, and an infrared (IR) emitter together with an IR depth sensor. By computing the IR depth data, the Kinect is able to detect and track up to two users based on either the distance of the subjects to the background or the subjects' body movement.

For each tracked user, it is able to robustly locate a total of 20 skeletal joints and track their movements over time in a 3D coordinate space (see Figure 4-1). It does not require any specific calibration posture or action from a user for tracking, while resulting in sufficient tracking accuracy. The tracking frame rate for a single user is able to reach about 30fps on i7 quad-core CPUs with 16GB RAM. For a comprehensive review of Kinect vision research, readers may refer to Han et al. (2013).

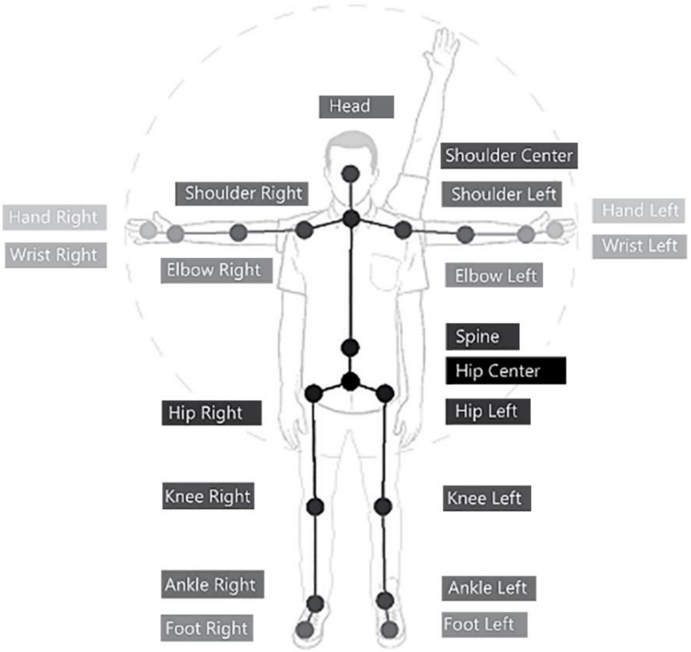


Figure 4-1 The tracked user’s skeleton with 20 joints, edited from Webb & Ashley (2012)

Our automatic affect recognition system is based on whole-body features extracted from the Kinect skeletal tracking data stream with 20 tracked joints in a geometric manner. Both the body form and movement information are modelled and extracted. The extraction of features is based on the recent psychology literature which indicates that some specific body behaviors may carry emotional information (e.g. Coulson, 2004; Harrigan et al., 2005). Studies in computer science automatically modeling affective body behaviors are also employed to guide the bodily feature extraction and selection (e.g. Kleinsmith et al., 2005; Kleinsmith et al., 2011; Savva et al., 2012; Metallinou et al., 2013). A comprehensive review on a variety of bodily expression features is provided in the work of Kleinsmith & Bianchi-Berthouze (2013). In this research, a total of 54 whole-body expression features (25 static posture features and 29 dynamic motion features) is extracted for each frame, and afterwards employed for the affective dimensional regression. These features range from lower-level features, such as the joint angles of elbow and knee, to more interpretable higher-level features, such as the lean angle of spine and the degree of body contraction/expansion. The comprehensive feature set and computation methods are summarized in Table 4-1.

These extracted features are potentially informative and may make distinctive contributions to each affective dimension. Furthermore, in order to identify the most optimal discriminative set of features for each dimension, we employ GA-based optimization to reduce feature dimensionality, as detailed in Section 4.1.2.

For some complicated features, we also provide the detailed explanations as below:

- **Body Expansion Index** measures the degree of contraction and expansion of the body, in frontal, lateral and vertical directions, respectively. Figuratively speaking, it computes a 3D bounding region, i.e., the minimum cuboid surrounding the entire body.
- **Instantaneous Velocity** can be calculated by dividing the displacement of a given joint between the current and last frames by the time interval of the two frames. It is related to the kinetic energy of a motion.
- **Average Velocity** states the averaged value of speed, and can be calculated by dividing the total motion trajectory length of a joint by the corresponding time interval.
- **Amplitude** indicates the maximum Euclidean Distance among the positions of a given joint within a predetermined time interval.
- **Acceleration** is the rate of change of velocity between the current and last frames. It is caused by the force applied to move the body part, and can be used to distinguish between smooth and sudden motions.

Table 4-1 Whole-body expression features and calculation methods

Feature Type	Features	Related Body Parts	Description & Calculation
Static Posture Features (25 in total)	**Body Expansion Index (in X, Y, Z axes)**	Whole Body	The degree of contraction or expansion of the whole body, in x, y and z axes
	Distance (in X, Y, Z axes)	Left hand to Left shoulder, Right hand to Right shoulder, Left hand to Left elbow, Right hand to Right elbow	The distance between the two given joints, in x, y and z axes (could be positive values, e.g. hand above shoulder; or negative values, e.g. hand below shoulder)
	Lean Angle	Head, Spine	The geometric angle of lean forward/backward
	Joint Angle	Left/Right elbows, Left/Right knees	The geometric angle of a given joint
	Euclidean Distance	Left hand - Right hand, Left elbow - Right elbow, Left hand - Right elbow, Right hand – Left elbow	The Euclidean distance between the two given joints
Dynamic Motion-based Features (29 in total)	**Instantaneous Velocity**	Head, Left/Right hands, Left/Right elbows	Instantaneous speed of a given joint, at the current frame
	Average Velocity (1s)	Head, Left/Right hands, Left/Right elbows	Average speed of a given joint within the past 1 second (≈30 frames)
	Average Velocity (3s)	Head, Left/Right hands, Left/Right elbows	Average speed of a given joint within the past 3 seconds (≈90 frames)
	Amplitude (1s)	Head, Left/Right hands, Left/Right elbows	Amplitude of a given joint within the past 1 second (≈30 frames)
	Amplitude (3s)	Head, Left/Right hands, Left/Right elbows	Amplitude of a given joint within the past 3 seconds (≈90 frames)
	Acceleration	Left/Right hands, Left/Right elbows	Instantaneous acceleration of a given joint, between two adjacent frames

4.1.2 Automatic feature selection based on GA optimization

Although great effort is spent on the feature extraction process, the 54 whole-body expression features listed in Table 4-1 are not necessarily of equal importance or quality. Some redundant or irrelevant features could result in an inaccurate conclusion whereas a compact and non-redundant subset of features could benefit subsequent regression models by improving their generalization and interpretability. Although domain knowledge could be applied to identify discriminative features, there is only limited understanding of how body posture and motion cues convey emotions due to the complexity of body language itself (Kleinsmith & Bianchi-Berthouze, 2013). Therefore, a GA-based automatic feature selection is employed to identify the most optimal discriminative feature subset for effective interpretation of bodily behaviors.

The GA, as a biologically inspired optimization search methodology based on a series of mechanisms mimicking Darwinian natural evolution and genetics in biological systems, is a promising alternative to conventional feature selection methods (Goldberg, 1989). The advantages of the GA for feature selection have been revealed by many studies (e.g. Oh et al., 2004; Huang & Wang, 2006). In a GA, a set of candidate solutions (called a population) to an optimization problem is evolved iteratively toward better solutions. In each iteration, each candidate solution (called an individual) is evaluated by a fitness function, and the more superior individuals are stochastically selected to form a new population (called a generation) through genetic crossover and mutation operation based on the Darwinian principle of ‘survival of the fittest’. The GA stops when the number of iterations reaches a preset threshold or acceptable results are obtained. Figure 4-2 illustrates a cycle of the GA evolutionary process. The details of our GA feature optimization are presented below.

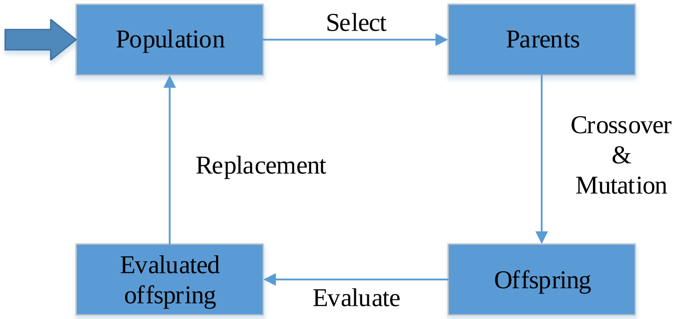


Figure 4-2 The evolutionary cycle of the GA

4.1.2.1 Chromosome encoding and population initialization

For the feature selection problem, solutions (i.e. selected features) are represented in a string with *n* binary digits, with each binary digit representing each feature, and values 1 and 0 meaning *selected* and *removed* features respectively. For example, chromosome '10001001' indicates the first, fifth and eighth features are selected. The GA starts with an initial population consisting of a number of *d* randomly generated solutions. In this research, the population size *d* is set to 20 according to original feature dimensions and computational complexity.

4.1.2.2 Fitness evaluation, selection, and replacement

The fitness evaluation for each chromosome normally consists of two criteria: prediction performance and number of selected features. Thus, the fitness function of a chromosome *C* is straightforward and defined as:

$$\text{fitness}(C) = w_a * \text{regression\_accuracy}_C + w_f * (\text{number\_features}_C)^{-1} \qquad (4\text{-}1)$$

where w_a and w_f are two predefined weights for regression accuracy and the number of selected features, respectively. Since the dimensions of the original dataset are relatively low (only 54), we focus on the regression accuracy rather than the number of selected features, i.e. the weight w_a is set to a large value (e.g. 0.9) whereas w_f is set to a much smaller value (e.g. 0.1).

During each successive generation, a proportion of the existing individuals is selected to form a new population for the next generation. According to Darwin's natural evolution theory, the fitter the individuals are, the higher the probabilities are to survive and create new offspring. Here, we adopt the roulette wheel selection mechanism (Goldberg, 1989). The probability that individual i is selected, $P(choice = i)$, is computed by:

$$P(choice = i) = \frac{\text{fitness}(i)}{\sum_{j=1}^{n} \text{fitness}(j)} \tag{4-2}$$

We then select two parent chromosomes based on the above method. The crossover operation subsequently generates two offspring out of the two parents, whereas the mutation operation slightly perturbs some offspring. The details of crossover and mutation are discussed in the following subsection. If the mutated offspring is fitter than both parents, the more similar parent is replaced by it; if it is fitter than only one parent, it replaces the inferior parent; otherwise, it replaces the most inferior individual in the population. We also employ an elitist selection strategy which allows some of the best individual solutions from the current generation to carry over to the next without alteration.

4.1.2.3 Genetic operation with crossover and mutation

The crossover and mutation functions are the two major factors that influence the fitness values of the generated individuals. We employ a standard crossover operator, i.e. single point crossover, for the exchange of genes between two parent chromosomes. Specifically, the binary string from beginning of chromosome to a random crossover point is copied from one parent, and the rest is copied from the other parent. The mutation mechanism is applied to the offspring, so that the genes may be altered occasionally. Specifically, in binary code, randomly selected bits are inverted, i.e. converting 0 to 1 or vice versa (see Figure 4-3). The newly generated offspring replaces the old population to form a new population in the next generation as discussed above.

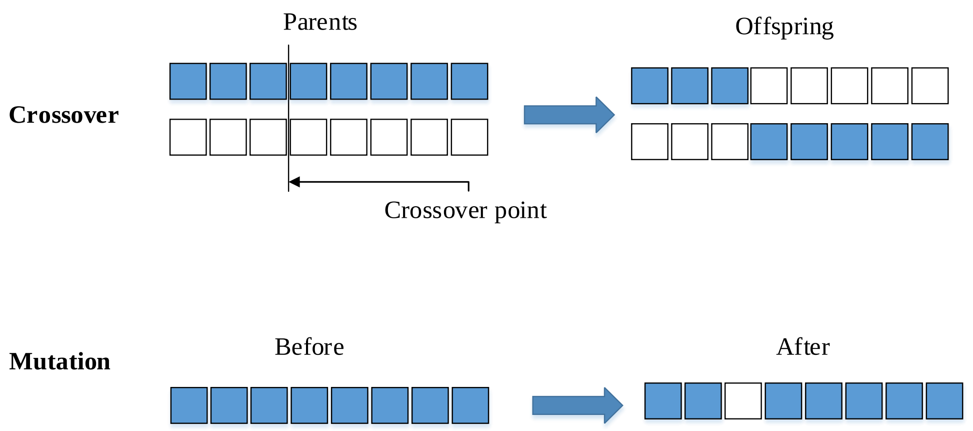


Figure 4-3 Examples of genetic crossover and mutation operations

4.1.2.4 GA Computational Complexity Analysis

Computational complexity normally refers to a property of a problem that how much computing resources are needed to solve the problem according to their intrinsic computational difficulty (Papadimitriou, 1994). It provides fundamental concepts for algorithm selection based on the rate of growth of space, time, or other fundamental unit of measure as a function of the size of the input (Bovet & Crescenzi, 1994).

According to Ankenbrandt (1991), GAs have a probabilistic convergence time. The average convergence time of a specific GA (typically measured as the number of generations to convergence) is possible to be determined by repeating an experiment a number of times. However, this average convergence may be mistaken for the complexity of the problem itself. Recent theory work (Rylander & Foster, 2000) and their follow-up study (Rylander & Foster, 2001) suggested that the GA-complexity can be measured by the growth rate of the minimum problem representation. Specifically, the GA-complexity of a problem is determined by the growth rate of the minimum representation as the size of the problem instance increases. In their work, a method based on Minimum Chromosome Length (MCL) was introduced predict the complexity of problems specific to GAs, which was then verified in two specific cases experimentally. These studies lead to the beginning of a theory that may enable us evaluate whether GAs are indeed efficient for a specific problem.

Soltani et al. (2002) suggest that GA performance can be measured by the number of fitness function evaluations carried out during the course of a GA run. In their work, a range of optimization algorithms (i.e. Dijkstra, A*, and GA) are compared and critically analyzed, and the GA is able to find the optimum or near-optimum solutions in considerably less execution time than the other two algorithms. They also indicate that efficiency of GA can also be analyzed by estimating the theoretical total number of possible solutions if an exhaustive search had been carried out, i.e. the size of the search space. This research is therefore motivated by Soltani et al. (2002) for the estimation and calculation of the computation efficiency of the GA.

The chromosome length equals to the number of features (i.e. 54) while the population size in each generation is set to 20, and the maximum generations is 2000. For fixed population sizes, the number of fitness function evaluations is given by the product of population size by the number of generations (Lobo et al., 2000). Thus, we can measure the computational complexity of GA as follows:

i. The theoretical total number of possible solutions to the problem, i.e. all possible combinations of features if we use a full enumeration search:

$$N_{total} = 2^{chromosome\ length} = 2^{54} \tag{4-3}$$

ii. The total number of 'actual' GA function evaluations:

$$\begin{aligned} N_{actual} &= (population\ size) * (number\ of\ generations) \\ &= 20 * 2000_{(maximum)} \end{aligned} \tag{4-4}$$

iii. Then, we calculate the ratio of: i. (total number of 'theoretical' possible solutions) to ii. (total number of 'actual' GA function evaluations):

$$Ratio = \frac{N_{total}}{N_{actual}} = \frac{2^{54}}{20*2000_{(maximum)}} \approx 4.5E + 11 \tag{4-5}$$

It can be noted from the ratio that the GA is able to generate the optimum or near-optimum solutions in substantially less execution time (i.e. $\frac{1}{4.5E+11}$) compared to that of a full enumeration search.

4.1.2.5 Parameter configurations

In this research, we apply the following parameter setting to achieve a balance between the regression accuracy and the computational complexity:

control procedure: steady-state;

population size = 20;

crossover probability = 1.0;

mutation probability = 0.05;

maximum generations = 2000;

These parameters are originated by the default setting of the GA algorithm with slight adjustment to our application domain which has an overall small feature set (i.e. 54 features). We perform GA-based optimization for both the arousal and valence dimensions, respectively. The selected feature subsets that lead to the best regression performance are finalized as the most discriminative subsets for each affective dimension, which is detailed in Section 4.3.2.

4.2 Dimensional affect interpretation using adaptive ensemble regression

To robustly predict the levels of affective dimensions (i.e. valence and arousal) from real-time bodily expression data stream, we propose an adaptive ensemble regression model that automatically generates and combines several base models to make a more reliable interpretation and regression of the valence and arousal dimensions. The proposed ensemble model is able to update itself and represents the most recent concepts in data streams. Therefore it has great adaptation to unseen bodily expression patterns and novel users. Feedforward Neural Networks with

Backpropagation (i.e. BPNNs) and Support Vector Machines for Regression (i.e. SVRs) are respectively used as the base regressors for the construction of the ensemble models for affective dimensional regression of bodily expressions. These techniques are also commonly used for continuous affect regression problems in the existing applications (e.g. Wollmer et al., 2008; Nicolaou et al., 2011). Experiments have also been conducted with single BPNN and SVR models for affective dimensional regression. The experimental results of such single regression models are also used as the benchmark for comparison.

Different from the ensemble classifiers proposed in Chapter 3.3, which aim to robustly differentiate between discrete emotions and detect novel emotion classes, the adaptive ensemble regression model proposed in this chapter is to effectively handle continuous affective dimension prediction tasks. Thus, we employ a series of different base models and ensemble mechanisms for the model generation, which are presented in detail below

Firstly, a number of bodily expression clips were collected from various participants for ensemble model generation and evaluation. Each clip collected in the dataset consists of a continuous sequence of instances (frames): $\{x_1, x_2, \ldots, x_i, y\}$, where x_i is an attribute (i.e. one of the bodily features listed in Table 4-1), and y is the target value (i.e. the annotated value of one affective dimension). The goal of a typical regression problem is to induce a function *f^(x)* on data consisting of a finite set of n instances to best approximate an unknown true function *f(x)*. In this research, we build an adaptive ensemble model that generates several base regressors that complement each other for robust regression of continuous affective dimensions.

The proposed adaptive ensemble regression model consists of two phases: **ensemble model generation** (during the training stage) and **regression and model updating** (during the test stage). Figure 4-4 illustrates the work flow of the generation of the ensemble model. The model generation phase starts with the weight initialization for each training clip, which is detailed in Section 4.2.1. Then a subset of training clips with higher weights is selected from the original training set. Subsequently, we train a base model using the newly generated training dataset with higher weights. Although a variety of algorithms, such as Decision Trees, could be used as the base regressor, in this research, we select BPNNs and SVRs respectively as the base regressors for the construction of two ensemble models. The details are discussed in Section 4.2.2. Subsequently, we calculate and assign a weight to the current base model based on its regression performance for the original training dataset. We also update the weights of the training clips with the aim of increasing the weights of those clips which have higher error rates and are more difficult to predict. The weight assignment and updating methods are detailed in Section 4.2.3. Overall, the above procedures iterate three times, thus three weighted base models are generated for ensemble, considering a balance between performance and computational complexity (Rokach, 2010). The final ensemble regression result can be thus obtained by calculating the weighted average of the outputs of the three base models.

Moreover, Figure 4-5 shows the flow chart of the automatic update of the ensemble model in the test stage. As mentioned above, the proposed ensemble model is able to deal with valence and arousal regression for newly arrived unseen bodily expression patterns to deal with data stream regression. In this research, such adaptability is achieved by gradually updating its base models with a stand-by base regressor. Once a new test instance arrives, it adds to the latest training dataset. The ensemble model generates a new stand-by base regressor using this new dataset. Then, we calculate and update the weights of both the newly generated and the original base models based on their prediction performance for the new dataset. If the new base model has a higher weight than any of the existing ones, then it is used to replace the base model with the lowest weight (i.e. the lowest regression accuracy). After that, an essential weight normalization procedure is performed for the updated base regressors. Thus, the ensemble model represents the latest concepts in the data and possesses great adaptation to the new data stream. The test stage of the ensemble model is discussed in

Section 4.2.4. For an exhaustive review of ensemble approaches for regression, readers may refer to Mendes-Moreira et al. (2012).

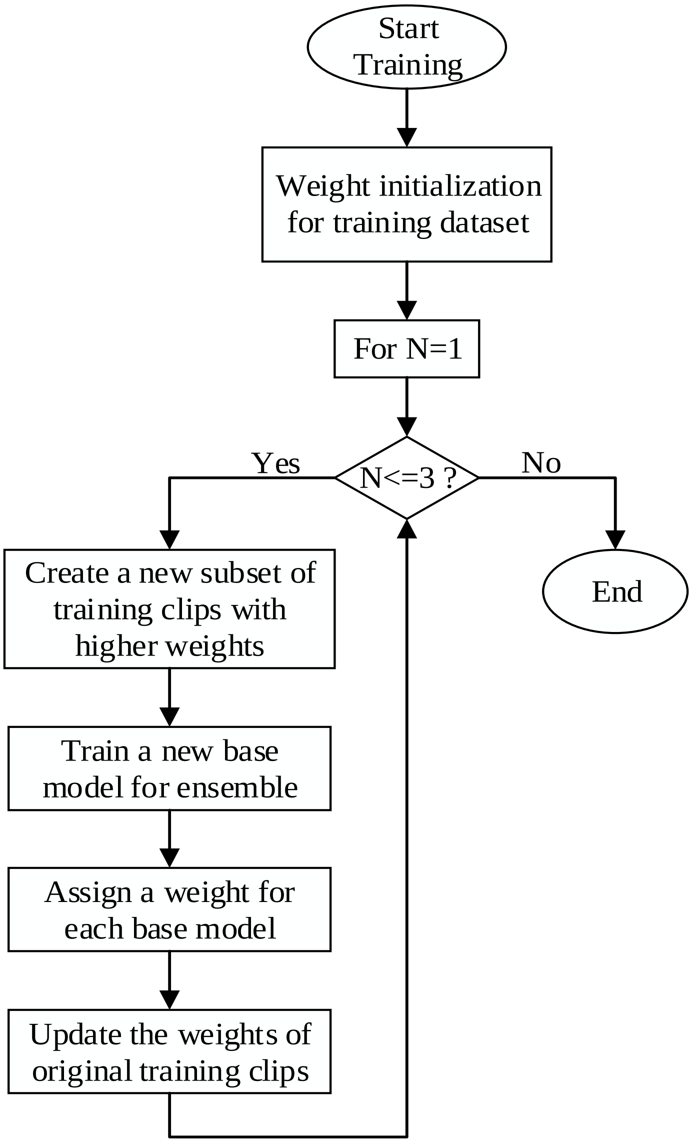

Start Training
Weight initialization for training dataset
For N=1
N<=3 ?
Yes
No
Create a new subset of training clips with higher weights
Train a new base model for ensemble
Assign a weight for each base model
Update the weights of original training clips
End

Figure 4-4 Flow chart for the generation of the proposed ensemble regression model

4.2.1 Weight initialization for training clips

First of all, we discuss the weight initialization of the training dataset. Many existing ensemble approaches (e.g. boosting algorithms) tend to initialize the weight of each training instance using an equal value. However, assigning appropriate weights has also been proved to increase the classification accuracy of the ensemble classifiers. For instance, Farid & Rahman (2013) assigned different weights for training instances based on the highest posterior probability generated by a Naive Bayes classifier, and demonstrated higher classification accuracies than uniform weight initialization.

In this research, we initialize a weight for each training clip based on the Pearson correlation coefficient (CORR) of a multiple linear regression analysis against the ground truth. That is, once a training clip is assigned a weight, the weight will be shared by all instances (frames) contained in that clip. A multiple linear regression model (David, 2009) linearly approximates the relationship between a set of i explanatory variables (x_1, x_2, ..., x_i) and the dependent variable y, which can be represented by the following equation:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_i x_{ij} \tag{4-6}$$

where β_0 denotes a constant value, and $\beta_1 - \beta_i$ are the coefficients of the explanatory variables. Inspection of the training dataset shows that the clips with higher weights (CORR) usually have greater agreement levels between annotators than others with lower weights.

4.2.2 Base model generation

Having initialized the weights for each training clip, the second step is to build a set of complementary and typical base models. In this research, the diversity and accuracy of each base model is achieved by manipulating the training data, i.e. we select a unique subset of training clips for the training of each base model. Specifically, in the first iteration, the base model is trained using a subset of clips that are initialized with higher weights and considered to be “more typical”, while in the following iterations, different subsets of clips that have higher regression errors in the valence and arousal space are selected to generate the corresponding base models. Thus, in each new iteration there are more “challenging clips” in the training subset.

As discussed earlier, in this research, we build two adaptive ensemble models, which respectively employ BPNN and SVR as their base regression models. BPNN and

SVR are chosen because they are among the most representative supervised algorithms for regression problems. Specifically, the former is a well-known adaptive algorithm (i.e. small changes in the training set may imply significantly different outputs), which is well suitable for ensemble learning models (Mendes-Moreira et al., 2012). The latter is usually regarded as a stable learning algorithm. However, it is sensitive to parameter and kernel function variations, and thus the model diversity can be also achieved through properly adjusting these parameters. Therefore, they are respectively employed for the construction of the two effective ensemble models for affective bodily behavior regression.

4.2.3 Weight calculation and update

The weight of each training clip is updated with the intention to increase the weights of those clips which are more challenging for affective dimensional regression in the valence and arousal space (i.e. with lower CORR). In order to update the weights appropriately, we firstly calculate the overall $CORR_{overall}$ for the original n training clips:

$$CORR_{overall} = \frac{1}{n}\sum_{i=1}^{n} CORR_i \tag{4-7}$$

Then, the updated weights can be calculated as follows:

$$w_{i,updated} = w_i * (\frac{1-CORR_i}{1-CORR_{overall}}) \tag{4-8}$$

where w_i is the original weight and $w_{i,\,updated}$ is the newly updated weight for clip i. Once the weights of all training clips are updated, the weights are normalized, so that their sum remains the same value as it was before. In this way, the weights of the instances that have higher regression accuracy are decreased while those of the instances that pose great challenges to affective dimensional regression are increased.

Furthermore, after a base model is generated, a weight is also assigned to this base model based on its prediction performance on the original training set, i.e. the CORR of the predictions obtained by a base model against the ground truth is assigned to this base model as its weight. Once the weights of the three base models are generated, then they are also normalized so that the sum equals to 1.

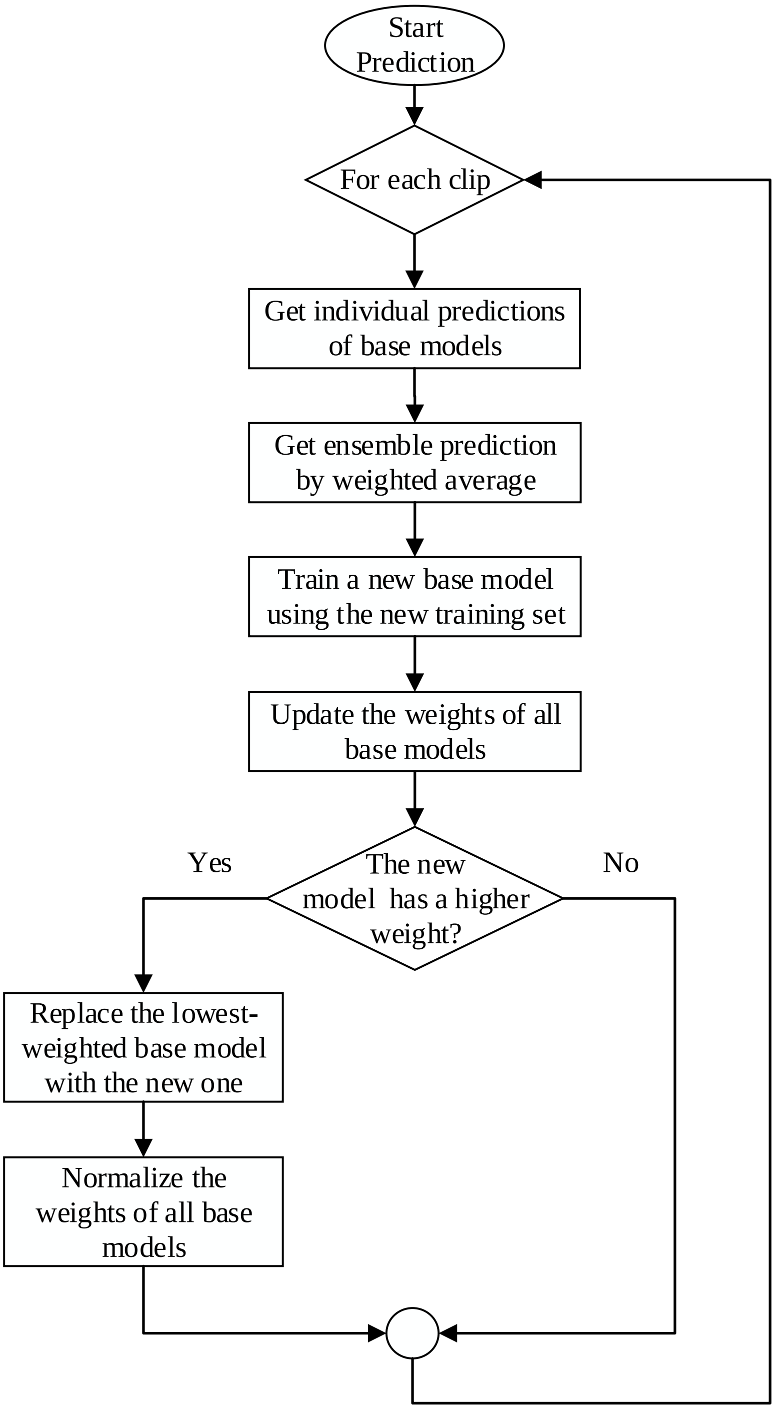

Start Prediction
For each clip
Get individual predictions of base models
Get ensemble prediction by weighted average
Train a new base model using the new training set
Update the weights of all base models
Yes
The new model has a higher weight?
No
Replace the lowest-weighted base model with the new one
Normalize the weights of all base models

Figure 4-5 Flow chart for ensemble regression and automatic model updating

By this stage, we have generated the proposed ensemble model. When dealing with dimensional regression of continuous bodily expression data streams in the test stage, the results of the base models are used to calculate the weighted average composite regression result as the final output for the affective behavior regression.

4.2.4 The adaptability of the ensemble regression model

Figure 4-5 presents the detailed steps for the generated ensemble model to deal with newly arrived test clips and automatic model updating. When a new test clip arrives, the update procedure starts with generating a new base model using the latest training dataset plus this newly arrived clip. Then we not only assign a weight to the new base model but also update the weights of the original base models based on their performance on this newly updated training dataset. For consistency, the CORR is used again to measure the prediction performance of each base regressor.

If the new base model has a higher weight than any of those existing models, then it is used to replace the minimum weighted base regressor. Thus, the new base model is added to the ensemble. Finally, the weights of the updated base models are normalized so that their sum remains ‘1’ (the same value as it was before).

4.3 Evaluation and discussion

In this section, we present the data collection, affective annotation and system evaluation for the proposed dimensional affect interpretation.

4.3.1 Data collection and dimensional affective annotation

First of all, we discuss whole-body expression data collection and affective annotation for evaluation. Since inappropriate fusion of annotations from multiple evaluators may significantly degrade the reliability and feasibility of an annotated corpus, we especially address the issue of continuous affective labelling and its follow-on problem of high inter-annotator disagreement.

4.3.1.1 Data collection of whole-body expressions

For our current study, eleven participants, five female and six male, ranging from 25 to 40 years old, were recruited for our bodily expression data collection. Most of them were postgraduate students and university lecturers, so that it is able to minimize the risks of inconsistent emotional expressions caused by differences between cultures, education backgrounds, and age groups, etc. Each participant was aroused to express various emotional states several times, and a total of 116 clips that contain more than 50,000 valid frames were recorded (detailed in the rest of this section). The number of participants and sample size are adequate compared to existing research (e.g. Nicolaou et al., 2011; Metallinou et al., 2013). Moreover, five annotators participated in total, rating overlapping for each participants, so that each recording would be annotated by five people justly (detailed in Section 4.3.1.2).

To ensure properly tracking participants' whole-body expressions (i.e. the 20 major joints illustrated in Figure 4-1), all participants were asked to stand in front of the Kinect with the distance between participants and the Kinect controlled within the range of 3 (±0.5) meters, so that it was able to achieve the best skeletal tracking effect. Before starting the data collection process, all participants were briefly trained, which allowed them to be more familiar and comfortable with the Kinect sensor and laboratory conditions to enable them to perform body language in a more natural way. Moreover, in order to avoid stereotypical and strongly acted expressions, we employed more diverse and interactive methods to arouse emotional responses of participants, such as viewing tragic/comedic movie clips, telling jokes, and making improvised performances with each other, instead of directly guiding them to perform specific emotional bodily expressions. Thus, it is able to well reflect the variety and subtleness of natural bodily expressions in real-life scenarios (e.g. high/low arousal/valence).

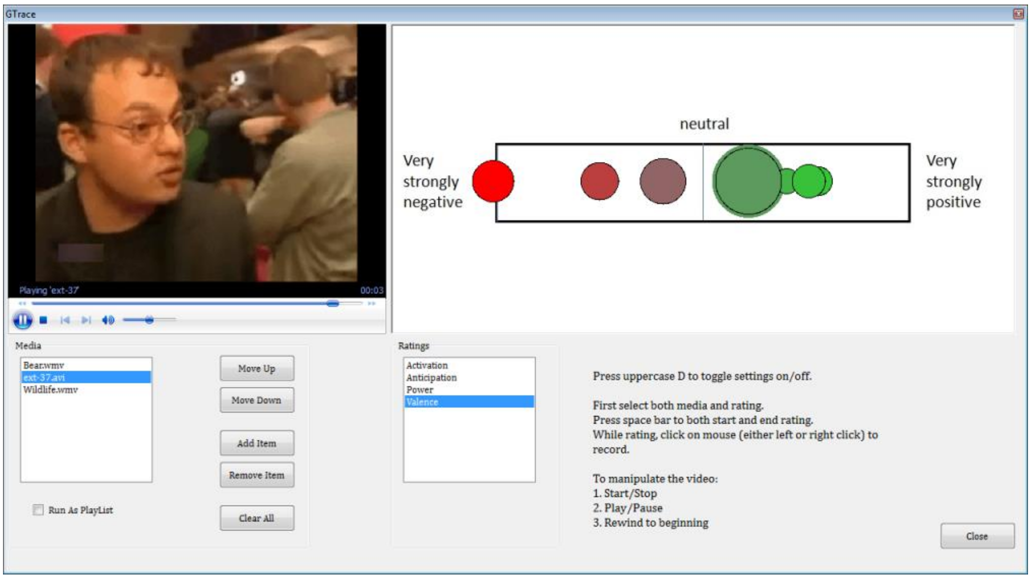
A total of 116 clips containing various emotional expressions was recorded (including both skeletal tracking data from the depth sensor and color video data from the RGB camera). The time length of each clip varies between 10 and 20 seconds (i.e. between approximate 300 and 600 frames per clip, with 450 frames on average, thus a total of more than 50,000 frames were collected). Each clip starts from a natural state and includes one or a few emotional bodily expressions (see Figure 4-7). After examining the skeletal tracing data, 31 out of the 116 clips were found to be considerably noisy (mainly due to involuntary sideways poses of participants, which may lead to tracking performance degradation since a part of the body is not visible to the sensor), and thus were excluded from this research. Therefore, our final bodily expression corpus contains 85 emotional bodily expression clips in total. Our modelling of emotions is based on both static body form and dynamic motion features extracted from the skeletal tracking data. The color videos have been used for our data collection and annotation in this work.

4.3.1.2 Continuous and dimensional affective annotation

In this experiment, the ground truth of emotions is established based on the analysis of observers' annotation rather than participants' self-statements. Because firstly, the self-statement about feelings may not be always consistent with their emotional behaviors (Kleinsmith & Bianchi-Berthouze, 2013). Furthermore, our automatic recognition system is built to model the observer's judgment rather the expresser. The building observer-based ground truth has been preliminarily addressed in several existing research applications. For example, Kleinsmith et al. (2011) measured agreement of annotators by iteratively comparing each pair of them. Meng et al. (2011) applied multi-labeling techniques that attempted to model the ranking of preferences instead of an absolute judgment, and thus can reduce the noise caused by a forced choice annotation approach.

However, the continuous and dimensional nature of the annotation task poses a great challenge in this research. It is difficult and not always possible to achieve high-level agreement between all participating annotators, even for expert annotators.

This is not only because it is considerably more difficult to achieve general agreement between annotators in rating the level of each affective dimension than discrete emotional categories, but also because continuous annotation itself requires more constant attention from observers. To address this issue, we present a systematic method to filter out noisy annotations and build reliable ground truth. The detailed method is presented as follows.



Press uppercase D to toggle settings on/off.

First select both media and rating.
Press space bar to both start and end rating.
While rating, click on mouse (either left or right click) to record.

To manipulate the video:
1. Start/Stop
2. Play/Pause
3. Rewind to beginning

Figure 4-6 Screenshot of the GTrace annotation tool

The whole period of each clip is continuously annotated frame-by-frame by five annotators, most of whom had essential experience in affective annotation tasks. All of them had to pass a short training session before starting the annotation work, where the definitions of the arousal and valence dimensions were explained, and the GTrace labelling tool (Cowie & Sawey, 2011) was introduced briefly. GTrace has been widely applied to emotion database annotation tasks, and it allows annotators to create real-time continuous annotations of participants' emotional states that appear to be changing over time (see Figure 4-6). The main interface of GTrace consists of video screen (top left), rating window (top right), and control panel (lower part of screen) which contains various selection options. Each annotator was first asked to view a number of clips to get an overall idea of our corpus, and then practice annotation with the first clip multiple times so that they can get more acquainted with the GTrace tool. During the annotation process, annotators were required to concentrate on only one dimension each round and encouraged to perform annotation as many times as desired for each clip until they felt satisfied with their annotations for each dimension. In this way, we are able to minimize person-specific instability during the real-time annotation tasks. Having obtained the annotations from each annotator, we subsequently focus on how to establish reliable ground truth for each affective dimension using these annotations.

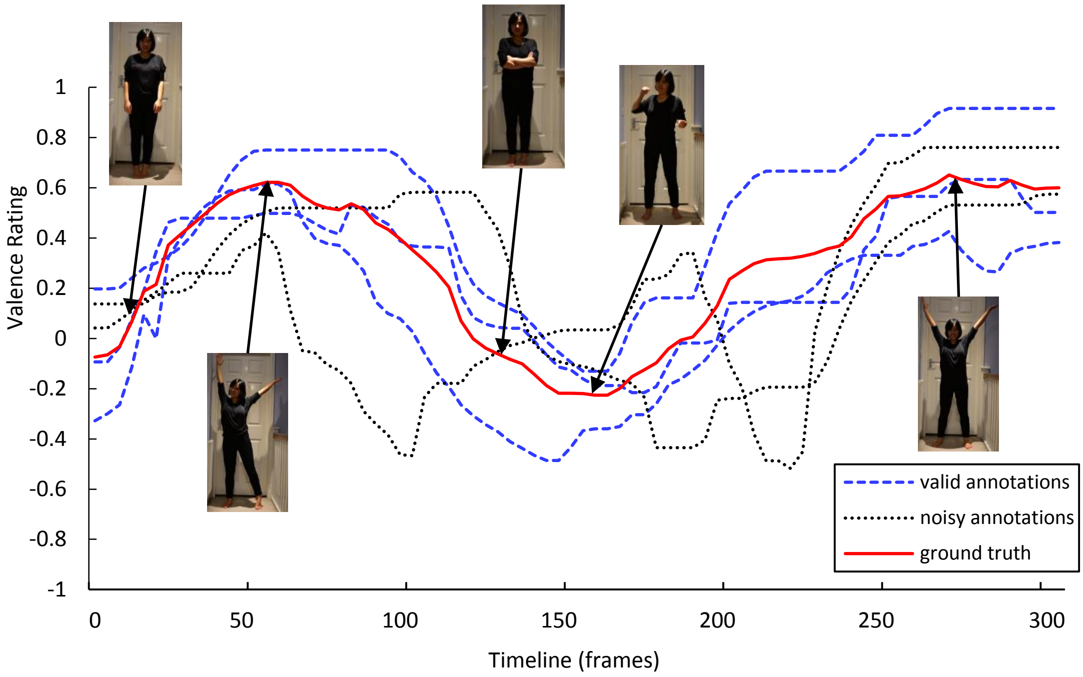


Figure 4-7 An example of valence rating for one clip by five annotators and the final calculated ground truth (The two grey dotted lines represent noisy annotations with CORR < 0.4)

We present an example segment of the valence annotations by the five annotators in Figure 4-7. The range of valence/arousal ratings is from -1 (the most negative/inactive) to +1 (the most positive/active) as mentioned above. As illustrated, the actual valence values from the five annotators could be different considerably at one time point. These differences are thought to be caused by inter-annotator variability (e.g. personal bias, annotation skill, and emotional state of annotators) and may typically occur in dimensional affective annotation tasks.

However, compared to the actual values, annotators tend to achieve higher-level agreement on the trends of the valence rating curves (i.e. general up-slope or down-slope). Such findings also hold truth for the arousal dimension and are consistent with previous research (e.g. Nicolaou et al., 2011; Metallinou et al., 2013). Thus, instead of using absolute values to evaluate the agreement levels between annotators, we determine to focus on the inter-annotator correlation, i.e. the Pearson correlation coefficient, to contribute to the ground truth generation.

For each clip, we apply the following three steps to establish the ground truth for both valence and arousal:

iv. We calculate the CORR for each pair of annotations, and then filter out the pair(s) with the CORR lower than a cutoff threshold (e.g. the two annotations show dramatic trend differences to each other as marked by the grey dotted lines shown in Figure 4-7).

v. We calculate the mean value of each annotation, and then filter out the pair(s) with the difference of the mean values greater than a cutoff threshold.

vi. Then the rest annotations are selected to compute the ground truth for the corresponding clip by taking the average of them. If there is no annotation left (i.e. all the five annotations are filtered out), that clip will be excluded from our corpora, as lacking essential inter-annotator agreement to establish the ground truth.

The cutoff thresholds for the CORR and the mean value difference are respectively set to 0.4 (a standard for moderate correlations in statistics) and 0.5, empirically. In this way, we select 68 and 72 valid emotional clips (with acceptable inter-annotator agreement and well-founded ground truth) for the valence and arousal dimensions respectively, out of the 85 clips produced in the previous step in total. The rest of the unselected clips will be excluded from further analysis as they could be either incomplete or ambiguous for emotional expression.

Although effectively fusing multiple annotations and generating proper ground truth based on different annotators' subjective judgments are challenging research problems (e.g. Audhkhasi & Narayanan (2013)), in this research, the use of the inter-annotator correlation and mean value difference metrics provides an effective solution for robust establishment of the underlying ground truth in continuous and dimensional annotation tasks. The presented method is able to effectively filter out potential noisy annotations (e.g. confusing or conflicting annotations with obvious trend differences or personal bias), and in the meanwhile it is more tolerant to non-noise value differences (e.g. different inter-annotator rating scales) that commonly exist in human affective annotation.

4.3.2 Experimental results and discussion

As mentioned earlier, a total of 72 (for arousal) and 68 (for valence) valid emotional clips from eleven participants is employed in our experiments, resulting in a rich corpus with around 45,000 samples (frames). All experiments are conducted following a leave-one-subject-out cross-validation scheme as it could be a more reliable evaluation method especially when the quantity of data/subjects is relatively limited. More specifically, we use the data of ten subjects for training, and the rest one for testing. This process is repeated 11 times (as we have eleven subjects in total), so that each subject can be tested in turn. The final cross-validation result is an average over these rounds.

In Table 4-2 and Table 4-3, we present the results of applying single regressors, i.e. BPNN and SVR, and the proposed ensemble models with BPNN and SVR as the base regressors respectively for the regression of arousal and valence dimensions using automatically selected features based on the GA optimization. The termination criteria of the GA optimization are that (1) the number of generations reaches 2000, or (2) the fitness value does not show obvious improvement during the last 50 generations. The best solution, i.e. the selected feature subset, is obtained when either termination criterion is satisfied. Since the GA optimization is a stochastic method, we perform a number of trials to find the most discriminative feature subset. Empirically, the GA is able to achieve convergence within 1000-1500 generations in most trials, and the number of selected features ranges between 25 and 40. The detailed results of each trial are presented in the first five rows of Table 4-2 and Table 4-3. We also perform three additional trials using manually devised features, i.e. either full set of static or dynamic features, or the combination of them. The results are presented in the last three rows.

First of all, as shown in experimental results, both BPNN-based and SVR-based ensemble models achieve consistently better performance than their corresponding single regressors for affective behavior regression in the arousal-valence space. More specifically, with the SVR-based ensemble models and the GA based feature optimization, we obtain the highest correlations with the ground truth (arousal: CORR=0.903, valence: CORR=0.815) and the lowest MSE values (arousal: MSE=0.057, valence: MSE=0.093) followed by the ensemble model with BPNNs as the base regressors which achieves comparable correlation (arousal: CORR=0.883, valence: CORR=0.811) and MSE (arousal: MSE=0.06, valence: MSE=0.105) values. These empirical findings indicate that the proposed adaptive ensemble models with the GA-based feature optimization are efficient and robust enough for challenging dimensional affect interpretation and regression tasks.

It is also hypothesized in this thesis that static and dynamic bodily features may contribute distinctively to different affective dimensions. We examine how the different combinations of features perform for the affect behavior interpretation and regression for the arousal and valence dimensions. With respect to the arousal dimension, it is observed in Table 4-2 that, in all the five trials of the GA-based feature optimization, the feature combinations selected consist of roughly equal numbers of static and dynamic features. The best regression results for the arousal dimension (CORR=0.903, MSE=0.057) are obtained using the optimal feature set generated by the GA with 19 static and 18 dynamic features and the SVR-based ensemble model. The best results of the BPNN-based ensemble and other single regression models are also achieved by employing the same feature set. Moreover, the sole use of static or dynamic features is still able to achieve relatively promising results (see the sixth and seventh trials in Table 4-2). These results suggest that both static and dynamic features play significant roles in the regression of the arousal dimension.

Compared to arousal, however, for the valence dimension, the feature combinations leading to promising results consist of the vast majority of static and only few dynamic features. As shown in Table 4-3, the best regression performance for valence (CORR=0.815, MSE=0.093) is achieved using the optimized feature set generated by the GA with 23 static and 2 dynamic features and SVR-based ensemble. It is also noticed that, by using static features exclusively (see the seventh trial in Table 4-3), we are also able to obtain relatively promising results. But on the contrary, the combination of entire sets of static and dynamic features does not provide any performance enhancement, although the results of solely using dynamic features show some basic positive correlations with the ground truth (see the sixth and eighth trials in Table 4-3). Inspection of clips with higher regression errors indicates that subjects with obviously different levels of valence can still have very similar patterns of bodily motion features in some cases (e.g. no matter if subjects are ecstatic or furious, they may unconsciously raise and shake their arms fiercely), and thus such dynamic features are considered to be less informative and may lead to confusion for the regression of valence. However, the role of dynamic features in valence prediction should not be dismissed entirely, as there is still great potential for further improvement by introducing more subtle and context-specific dynamic features.

Moreover, by the comparison between Table 4-2 and Table 4-3, it indicates that the arousal dimension regression performance generally outperforms the valence dimension. This suggests that the bodily expressions could be a better indicator of the arousal dimension than valence. This result is also supported theoretically by Ekman & Friesen (1967), and largely consistent with recent research of continuous affect

Table 4-2 Regression performance for arousal using the GA-based feature optimization (the first five trials) and manually devised features (the last three trials)

Trials.	**Number of selected static features**	**Number of selected dynamic features**	**BPNN**		**SVR**		**Ensemble (NN)**		**Ensemble (SVR)**	
			CORR	**MSE**	**CORR**	**MSE**	**CORR**	**MSE**	**CORR**	**MSE**
1	18	16	0.778	0.076	0.845	0.069	0.856	0.068	0.885	0.064
2	19	17	0.784	0.073	0.855	0.064	0.864	0.066	0.891	0.061
3	17	19	0.795	0.072	0.864	0.062	0.871	0.063	0.902	0.061
4	19	18	**0.797**	**0.069**	**0.876**	**0.061**	**0.883**	**0.06**	**0.903**	**0.057**
5	18	22	0.778	0.076	0.845	0.069	0.856	0.068	0.885	0.062
6	/	29 (entire set)	0.687	0.11	0.726	0.108	0.728	0.111	0.745	0.102
7	25 (entire set)	/	0.733	0.102	0.796	0.095	0.807	0.094	0.81	0.091
8	25 (entire set)	29 (entire set)	0.776	0.077	0.845	0.072	0.853	0.069	0.882	0.067

Table 4-3 Regression performance for valence using the GA-based feature optimization (the first five trials) and manually devised features (the last three trials)

Trials.	**Number of selected static features**	**Number of selected dynamic features**	**BPNN**		**SVR**		**Ensemble (NN)**		**Ensemble (SVR)**	
			CORR	**MSE**	**CORR**	**MSE**	**CORR**	**MSE**	**CORR**	**MSE**
1	21	4	0.73	0128	0.783	0.11	0.803	**0.105**	0.809	0.102
2	23	2	**0.733**	**0.127**	**0.792**	**0.105**	**0.811**	0.111	**0.815**	**0.093**
3	22	3	0.726	0.13	0.769	0.106	0.783	0.123	0.791	0.111
4	21	5	0.722	0.135	0.781	0.116	0.794	0.115	0.788	0.107
5	20	6	0.727	0.141	0.762	0.121	0.788	0.119	0.793	0.102
6	/	29 (entire set)	0.374	0.239	0.412	0.226	0.789	0.217	0.42	0.199
7	25 (entire set)	/	0.728	0.129	0.781	0.115	0.772	0.113	0.798	0.095
8	25 (entire set)	29 (entire set)	0.719	0.131	0.766	0.117	0.767	0.104	0.78	0.112

modelling (e.g. Kleinsmith et al., 2011; Metallinou et al., 2013) which claimed bodily expressions tend to convey less information about valence in comparison with other affective dimensions. Furthermore, there is evidence that the valence dimension could be better reflected and recognized by other modalities, such as facial expressions (Kleinsmith & Bianchi-Berthouze, 2013). Thus, it could be quite promising to explore the fusion of whole-body expressions with facial expression detection, which will be detailed in Chapter 5.

The proposed system has been also applied to real-time bodily affect regression tasks. The computational complexity of the skeletal tracking normally needs 10-15 milliseconds. The feature extraction, selection and affect behavior ensemble regression require an averaged run time of 3-10 milliseconds (which may vary with different base models used). Overall, this system is able to perform efficiently and reach around 30fps on i7 quad-core CPUs with 16GB RAM.

4.3.3 Comparison with state-of-the-art performance

Furthermore, we compare the proposed system with other state-of-the-art developments. These methods (as listed in Table 4-4) are selected as the benchmarks because they produced state-of-the-art performance using similar bodily features for continuous and dimensional affect interpretation and regression, and presented their results through the same metric (i.e. the CORR). As shown in Table 4-4, as both Nicolaou et al. (2011) and Metallinou et al. (2013) applied various modalities and leaning models, we only use their best results for a more explicit comparison.

In Nicolaou et al. (2011), the bodily features employed in their work only contained static shoulder points, which are much simpler compared to our feature sets. Their work also obtained a comparably promising result for valence (CORR=0.796). It could be attributed to the fact that they also employed other modalities (e.g. facial expressions) to incorporate with their bodily features, which are able to greatly boost the prediction performance for valence. Moreover, Metallinou et al. (2013) employed full-body language features and Gaussian Mixture Model (GMM)-based approach to track continuous levels of valance, arousal and dominance in inter-personal interactions. Their system achieved a relatively lower performance (CORR=0.584 and 0.225, for arousal and valence, respectively). It may be attributed to inadequate features employed, e.g. the dynamic features they extracted were only concerned with velocity. Overall, in comparison with related research, our system consistently outperforms the above applications reported in the literature. The well-refined whole-body features and the proposed SVM-based adaptive ensemble model enable us to achieve the best regression performance for both arousal (CORR=0.903) and valence (CORR=0.815) dimensions. Overall, the above comparison further proves the effectiveness of our proposed system for continuous and dimensional affect regression.

Table 4-4 Comparison with related research

(SAL: SAL database (Douglas-Cowie et al., 2007), BLSTM-NN: Bidirectional Long Short-Term Memory neural network, LSTM: Long Short-Term Memory neural network, GMM: Gaussian Mixture Model)

	Feature type	Static (posture) / Dynamic (motion)	Learning Model	Database/Number of sample used	Performance (CORR)	
					Arousal	Valence
Nicolaou et al., (2011)	Shoulder point / facial / audio features	Static (with temporal information)	SVR; BLSTM-NN	SAL, 4 subjects, 30,000 visual and 60,000 audio samples	0.642	0.796
Metallinou et al., (2013)	Body language / audio features	Static & Dynamic (partially)	LSTM-NN; GMM model	Private dataset, 16 subjects, 100 recordings	0.584	0.225
This work	Whole-body expression features	Static & Dynamic	BPNN; SVR; Ensemble model	Private dataset, 11 subjects, 140 clips, 45,000 samples (frames)	**0.903**	**0.815**

4.4 Summary

In this chapter, we shed light on the bodily modality and address the problem of continuous affective dimensional regression using whole-body expressions in an arousal-valence space. We systematically extract both users' static posture and dynamic motion-based bodily features. The GA is subsequently applied to perform feature optimization to identify their optimal discriminative combination for the regression of each dimension. We also propose an adaptive ensemble regression model to robustly predict affective dimensions and map users' affective states into an arousal-valence dimensional space. The proposed adaptive ensemble model employs the weighted average for regression and significantly outperforms other single model based methods, in terms of both regression accuracy (MSE) and correlation (CORR). It also shows good adaptation to newly arrived unseen bodily expressions.

Our empirical findings also indicate that static and dynamic bodily features have distinctive contributions to different affective dimensions, especially to valence. Specifically, the combination of static posture and dynamic motion features achieves the best regression performance for arousal, whereas the static posture features seem to contribute more than dynamic features for the regression of valence. Also, arousal is generally better predicted than valence in this research, which is also consistent with both psychological literature (e.g. Ekman & Friesen, 1967) and other dimensional affect recognition research (e.g. Nicolaou et al., 2011; Metallinou et al., 2013). Overall, the proposed system with the SVM-based ensemble model outperforms existing research reported in the literature and achieves the best regression performance for both arousal with CORR=0.903 and MSE=0.057, and valence with CORR=0.815 and MSE=0.093 respectively with a promising real-time performance of 30fps.

Chapter 5 Fusion of facial and bodily modalities for enhancing dimensional affect interpretation

In this chapter, we present a bimodal dimensional affect recognition system by incorporating affective information from both bodily and facial modalities. We propose a semi-feature level fusion framework that integrates users' whole-body expression features with facial Action Unit intensities and demonstrate significantly improved regression prediction performance for dimensional affect interpretation. Section 5.1 reviews the state-of-the-art developments in bimodal/multimodal emotion recognition. In Section 5.2, we present the detailed methodology of the proposed semi-feature level fusion. Experiments, evaluation and discussion are presented in Section 5.3.

5.1 Review of state-of-the-art developments

Automatic emotion recognition is a well-established and fast growing field, and there is an extensive literature available on emotion recognition from different modalities (or their combinations). It has been widely acknowledged that the use of multimodal information allows for a more complete emotional description and enables more accurate recognition results. Currently, the mainstream multimodal research has mostly focused on the recognition of facial and vocal expressions in terms of a small number of discrete emotion categories (e.g. Gunes et al., 2008; Gunes & Pantic, 2009; Cohn et al., 2009). For an extensive survey on multimodal emotion recognition research, readers may refer to Zeng et al. (2009).

Table 5-1 Summary of multimodal and dimensional affect recognition systems (SAL: SAL database (Douglas-Cowie et al., 2007), BLSTM-NN: Bidirectional Long Short-Term Memory Neural Network, LSTM: Long Short-Term Memory Neural Network, BPNN: Backpropagation Neural Network, LDA: Linear Discriminant Analysis, SVM: Support Vector Machine, SVR: Support Vector Regression, GMM: Gaussian Mixture Model)

System	Modality/Feature type	Database/Number of sample	Learning/Classification model	Fusion strategy	Results
Karpouzis et al. (2007)	Various visual & acoustic features	SAL, 4 subjects, 76 passages	Recurrent Network with 4 class-outputs	not reported	Negative/positive/active/passive, 67% recognition accuracy with vision, 73% with prosody, 82% after fusion
Kim (2007)	Speech & physiological signals	Private database, 3 subjects, 343 samples	Modality-specific LDA-based classification	Integration of feature and model-level fusion	4 Arousal-Valence quadrants, 55% for feature fusion, 52% for decision fusion, 54% for hybrid fusion
Nicolaou et al. (2010)	Facial expression, shoulder gesture, audio cues	SAL, 4 subjects, 30,000 visual and 60,000 audio samples	HMM and likelihood space via SVM	Model-level fusion, likelihood space fusion	Negative vs. positive valence (quantized), 91.76% by facial expressions, 94% by modal fusion
Nicolaou et al. (2011)	Facial expression, shoulder gesture, audio cues	SAL, 4 subjects, 30,000 visual and 60,000 audio samples	SVR and BLSTM-NN	Feature/model-level, output-associative fusion	Valence and arousal (continuous), best results: RMSE=0.15 and CORR=0.796 for valence; RMSE=0.21 and CORR=0.642 for arousal
Metallinou et al. (2013)	Body language and speech cues	Private database, 16 subjects, 100 recordings	LSTM and GMM-based prediction	Feature-level fusion	Valence, arousal and dominance (continuous), CORR=0.584, 0.056, 0.337, respectively
This work	Facial and whole-body expressions	Private database, 11 subjects, 40,000 samples (frames)	BPNN, SVR, and proposed ensemble models	Semi-feature level fusion	Valence and arousal (continuous), MSE= 0.077 and CORR= 0.886 for valence; MSE= 0.056 and CORR= 0.907 for arousal

It has been shown that in real-life interactions people tend to exhibit more subtle and complex emotional states rather than only a small number of basic discrete emotion categories acquired in laboratory settings. This poses a great challenge to the aforementioned systems which aim to describe users' emotional state by single discrete labels. Thus, it is not surprising that a growing body of research has recently focused on dimensional affect recognition. For example, Karpouzis et al. (2007) employed a Simple Recurrent Network which lends itself well to modeling dynamic events in both users' facial expressions and speech for the recognition of emotion in naturalistic video sequences. In their work, a quantized dimensional representation of users' emotional states (i.e. activation and valence) was applied, instead of detecting discrete emotion categories. Kanluan et al. (2008) employed late fusion of facial expression and audio channels by using weighted linear combinations of their outputs respectively obtained by SVM for regression to estimate the valence, activation, and dominance dimensions (on a 5-point scale, for each dimension).

Most recently, a few attempts have been proposed for actual continuous affective dimension regression (without quantization). For example, Nicolaou et al. (2011) employed three modalities including facial expression, shoulder gesture and vocal cues for continuous tracking of the valence and arousal affective dimensions. Metallinou et al. (2013) proposed a Gaussian Mixture Model-based approach to continuously predict levels of participants' activation, valence and dominance during the course of affective dynamic interactions using body language and speech features. For a more clear comparison, in Table 5-1, we briefly summarize some state-of-the-art applications that employ multiple modalities to model and recognize affect in terms of affective dimensional space, together with our work presented in this chapter. Although some earlier applications listed in Table 5-1 (Karpouzis et al., 2007; Kim, 2007; Nicolaou et al., 2010) applied a discretized classification scheme rather than a continuous dimensional space, we still include them as they are relevant to this study.

In comparison to the existing work listed in Table 5-1, our research presents the first semi-feature level fusion framework in the literature that effectively combines users' whole-body features and facial Action Unit intensities to improve prediction performance for affective dimensions. The detailed fusion method is presented in the following.

5.2 Modality fusion strategy for dimensional affect interpretation

As illustrated in Figure 5-1, the proposed semi-feature level fusion is realized by concatenating the derived AU intensities (as discussed in Section 3.2) and the optimal discriminative bodily features (as discussed in Section 4.1) into a new feature vector which is subsequently employed as inputs to affective dimensional regressors. A feature normalization procedure is also performed, in which each attribute is linearly scaled to the range of [0; +1]. The adaptive ensemble regression model proposed in Section 4.2 is employed for our bimodal affective interpretation as it outperforms the two other benchmark single models, i.e. BPNN and SVR.

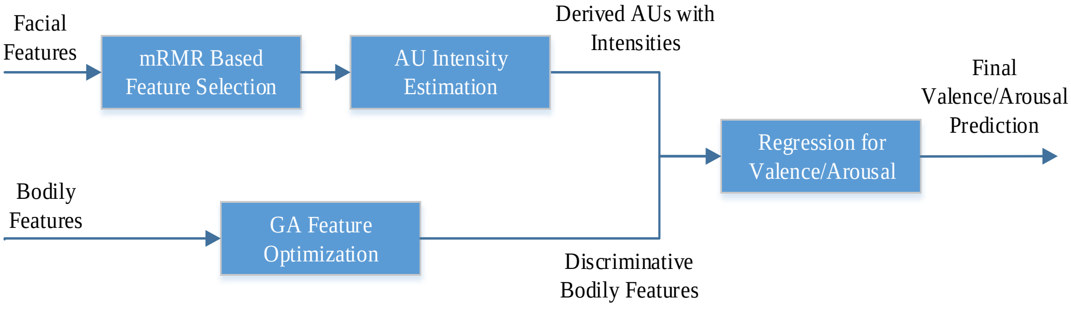

Facial Features
mRMR Based Feature Selection
AU Intensity Estimation
Derived AUs with Intensities
Final Valence/Arousal Prediction
Regression for Valence/Arousal
Bodily Features
GA Feature Optimization
Discriminative Bodily Features

Figure 5-1 The proposed semi-feature level fusion framework

Our motivation is threefold. Firstly, there is strong psychological evidence (e.g. Ekman & Friesen, 1967; Ekman & Friesen, 1983) indicating that the bodily expressions could be a better indicator of the arousal dimension, whereas some facial actions convey rich information of the valence dimension (e.g. the occurrence of AU1 Inner Brow Raiser usually indicates a 'sad' emotion, whereas AU12 Lip Corner Puller normally occurs with ‘happiness’). Thus, their combination is able to contribute more complementary information for dimensional affect prediction.

Secondly, in this chapter we focus on dimensional interpretation of affect. Because in such an approach, even complex/blended emotion expressions and subtle emotion transitions can be captured and represented properly using continuous scale of different dimensions, which could be too difficult to deal with through the categorical approach.

Most importantly, although it remains largely unclear how humans achieve effective fusion of multimodal affective signals for a final decision, recent literature (Stein & Meredith, 1993; Zeng et al., 2009) was more supportive of early stage fusion (e.g. feature-level fusion) rather than late stage fusion (e.g. decision-level fusion), because the feature-level fusion is able to catch more information and relations of different modalities to inform affect interpretation. However, it is difficult to directly combine features from different modalities with various metrics, dimensionalities and temporal structures. Thus, we propose the semi-feature level fusion that appropriately integrates the derived AU intensities with GA-optimized discriminative bodily features for dimensional affective interpretation, which is evaluated in the following section.

5.3 Evaluation and discussion

Our established corpus (as discussed in Chapter 4) with 85 emotional clips across eleven subjects used for the previous bodily affect recognition is employed for the evaluation of the proposed bimodal affect recognition with semi-feature level fusion. We select a total of 60 (for arousal) and 58 (for valence) valid emotional clips out of the 85 clips that has both effective skeleton and facial landmark tracking data for our experiments. We also follow a leave-one-subject-out cross-validation scheme, i.e. the data of ten subjects are used for training and the rest one for testing, and each subject is tested in turn. The final result is an average over these rounds. As mentioned earlier, the merged feature vector consists of the derived AU intensities and the most discriminative bodily features. The AU intensities are obtained from the SVR-based AU intensity regressors which are pre-trained with database images, whereas the bodily features are selected based on the GA optimization.

In Table 5-2, we present the experimental results of applying the ensemble regression models with BPNNs and SVRs as the base regressors respectively for arousal and valence dimensions using the merged feature vector by semi-feature level fusion. As shown in Table 5-2, the fusion of facial and bodily modalities provides obvious performance enhancement for both arousal and valence dimensions. Especially for valence, integrating facial AU intensity information appears to perform much better than solely using bodily features in terms of both MSE (0.077 vs. 0.093) and CORR (0.886 vs. 0.815). These results are theoretically consistent with psychological research (e.g. Ekman & Friesen, 1967; Ekman & Friesen, 1983) which hypothesizes that facial expressions communicate rich and explicit affective information of the valence dimension (e.g. happiness and sadness). These results demonstrate that the proposed semi-feature level fusion framework provides an effective solution for facial and bodily modality fusion, and achieves very promising performance improvements.

Table 5-2 Experimental results of the proposed semi-feature level fusion for arousal and valence

	Modality	**Ensemble (NN)**		**Ensemble (SVR)**	
		CORR	**MSE**	**CORR**	**MSE**
Arousal	Bodily	0.883	0.06	0.903	0.057
	Bimodal	**0.889**	**0.058**	**0.907**	**0.056**
Valence	Bodily	0.811	0.111	0.815	0.093
	Bimodal	**0.872**	**0.083**	**0.886**	**0.077**

5.4 Summary

In this chapter, we propose a semi-feature level fusion framework that incorporates affective information of both the facial and bodily modalities to draw a more reliable interpretation of users’ emotional states. Experimental results show that the proposed adaptive ensemble regression model achieves remarkable performance improvements for the regression of both the arousal and valence dimensions by combining the optimal discriminative bodily features and the derived AU intensities as inputs, in comparison to solely applying the bodily features.

Chapter 6 Conclusion and future work

In this research, we focused on automatic affect recognition based on facial and bodily modalities. In this chapter we summarize the principle contributions arising from our work and then identify potential future directions.

6.1 Summary of contributions

A number of core contributions have been raised in the research presented in this thesis. First of all, we proposed two different types of adaptive ensemble models (i.e. ensemble classification with novel emotion class detection and ensemble regression with adaptability to newly arrived unseen patterns), which are respectively tailored to discrete facial expression recognition and continuous dimensional bodily emotion regression tasks. We also made efforts in the stage of feature extraction and selection. An mRMR-based method and the GA optimization are employed for automatic feature selection from facial and bodily expressions respectively. The empirical findings indicate that these feature selection processes benefit the subsequent emotion recognition and regression significantly. Furthermore, a semi-feature level fusion framework has been also proposed to effectively integrate affective information from both the facial and bodily modalities for a more reliable and comprehensive emotion interpretation. We discuss these contributions in more detail below.

6.1.1 Facial action intensity regression and categorical emotion recognition

For AU intensity estimation, we employed dynamic motion-based facial features (e.g. the elongation of mouth) rather than static features (e.g. the width of mouth) as in many previous literatures. The motion-based facial features are caused by underlying facial muscle movements and thus are relatively universal and subject-independent for the expression of the six basic emotions, whereas the static features could change a lot between different subjects. These motion-based features were subsequently selected by using both manual and mRMR-based automatic methods, and then employed as inputs to 16 Neural Networks and Support Vector Regressors for AU intensity estimation, with each regressor dedicated to each diagnostic AUs. The mRMR-based automatic feature selection achieved comparable performance in comparison to the manually well-devised features.

We also proposed a set of six adaptive ensemble classifiers to differentiate between the six basic emotions and identify newly arrived unseen novel emotions using the derived AU intensities. Each ensemble classifier employs a special type of Neural Network, i.e. Complementary Neural Network, as the base classifier, which is able to provide uncertainty measure of its classification performance. The uncertainty measures and a distance-based clustering are used to inform the arrival of novel unseen emotion classes. Both off-line and on-line evaluation results demonstrated that the ensemble classifiers have great robustness and flexibility for not only the recognition of six basic emotions but also the detection of newly arrived unseen novel emotions.

6.1.2 Dimensional emotion regression based on whole-body expressions

In order to robustly map subjects' affective bodily expressions onto a valence–arousal space, we systematically extracted both static and dynamic whole-body features and applied the GA optimization to conduct feature selection and identify their optimal discriminative combinations for the regression of each affective dimensions. We also proposed an ensemble regression model with great adaptability for the regression of each dimension, which also employs a stand-by regressor to better deal with newly arrived unseen bodily expressions and novel subjects. Our empirical findings first proved that static and dynamic bodily features have distinctive contributions to different dimensions, e.g. static posture features seem to contribute more significantly than dynamic features for the valence dimension.

Moreover, the high level of disagreement between different annotators is inherently a problem of continuous and dimensional affective annotation. We also presented a novel annotation method that takes consideration of both the correlation between different annotators and the personal bias metrics to build reliable ground truth for system evaluation.

6.1.3 Bimodal emotion regression using semi-feature level fusion

There is recently a shift of focus from discrete and unimodal emotion recognition to continuous and multimodal recognition, as the latter is more flexible and reliable for the interpretation of spontaneous emotions in real-life scenarios. Thus, we also proposed a bimodal dimensional affect recognition system by semi-feature level fusion of facial and bodily modalities and achieved significantly performance improvements for the regression of both arousal and valence. To the best of our knowledge, this is the first attempt to combine AU intensities and whole-body features for automatic affect recognition, and overcomes the inherent shortcomings of conventional feature and decision-level fusion.

6.2 Future work

In this section, we identify the following several potential directions for further work. First of all, although we have employed different public databases and privately collected data for system evaluation, these data are all recorded under laboratory conditions. As pointed out by Kleinsmith & Bianchi-Berthouze (2013), a more naturalistic and extensive corpus with various subjects and challenging spontaneous affective expressions could better reflect the system performance. Thus, we will further validate the system’s performance in more challenging real-life interaction scenarios, since in spontaneous emotional expressions, both AUs and bodily expressions usually occur with relatively lower intensities in more subtle combinations comparing to the posed ones. Besides, by using an extensive database (i.e. with a larger number of subjects or labeled in a richer affective space with other dimensions, such as dominance and expectation), the proposed arousal-valence dimensional emotion recognition framework can be easily extended to other dimensions, and we can also further explore the correlations between those different affective dimensions.

Furthermore, literature indicates that, in some cases, the performance of ensembles could be potentially boosted by combining different types of base learning algorithms in one ensemble (Mendes-Moreira et al., 2012). Thus, it shows potential to further improve the proposed adaptive ensemble models by exploring such combinations of diverse base models. Moreover, although the GA-based feature selection shows advantages compared to other deterministic algorithms, in the future it would be further improved in various ways, such as exploring more suitable genetic operators or gene rearrangement algorithm for chromosomal encoding. Besides, further tuning of genetic parameters, such as analyzing the effect of different population sizes may also leave some room for further improvement.

Finally, the proposed adaptive ensemble emotion recognition systems could be integrated in various real-life applications and the benefits are evident in many areas of society, such as security surveillance, health care, interactive entertainment, and education. For example, students may lose motivation and efficiency when high levels of negative emotional states such as anxiety, frustration, and fear of failure are experienced (Kapoor et al., 2007). A computer-assisted learning system is able to read affective states of students from their facial expression and body language and react appropriately (e.g. adjust course difficulty and teaching speed) in an effort to help students maintain adequate motivation and efficiency. We believe that in the near future, the proposed systems may play an important role in our daily life.

References List

Ahlberg, J. (2001). CANDIDE-3—an updated parameterized face. *Report No. LiTH-ISY-R-2326*, Department of Electrical Engineer-ing, Linkoping University, Sweden.

Ankenbrandt, C. (1991). An extension to the theory of convergence and a proof of the time complexity of Genetic Algorithms. *Foundations of Genetic Algorithms*, Morgan Kaufman.

Atkinson, A.P., Dittrich, W.H., Gemmell, A.J., & Young, A.W. (2004). Emotion perception from dynamic and static body expressions in point-light and full light displays. *Perception*, 33 (6), 717–746.

Atkinson, A.P., Tunstall, M.L., & Dittrich, W.H. (2007). Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures. *Cognition*, 104 (1), 59–72.

Audhkhasi, K. & Narayanan, S. (2013). A globally-variant locally constant model for fusion of labels from multiple diverse experts without using reference labels. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(4), 769–83.

Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., & Movellan, J.R. (2006). Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1 (6), 22–35.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing - Letters and Review*, 11 (10).

Bernhardt, D. (2010). Emotion inference from human body motion. *Technical reports published by the University of Cambridge, Computer Laboratory*, ISSN 1476–2986.

Bernhardt, D., & Robinson, P. (2007). Detecting affect from non-stylized body motions.

LNCS: Process *2nd Int. Conference on Affective Computing and Intelligent Interaction*, 59–70.

Besinger, A., Sztynda, T., Lal, S., Duthoit, C., Agbinya, J., Jap, B., Eager, D., & Dissanayake, G. (2010). Optical flow based analyses to detect emotion from human facial image data. *Expert Systems with Applications*, 37, 8897–8902.

Bianchi-Berthouze, N. & Kleinsmith, A. (2003). A categorical approach to affective gesture recognition. *Cognitive Science*. vol. 15, 259–269.

Boone, T.R., & Cunningham, J.G. (2001). Children’s expression of emotional meaning in music through expressive body movement. *Journal of Nonverbal Behavior*, 25 (1), 21–41.

Bouguet, Y. L. (1999). Pyramidal implementation of the Lucas–Kanade feature tracker. *Technical Report*, Intel Corporation, Microprocessor Research Labs.

Bovet, D., & Crescenzi, P. (1994). *Computational Complexity*. Prentice Hall.

Bull, P.E. (1987). Posture and Gesture, volume 16. Pergamon Press.

Camurri, A., ILagerl ̈of, I., & Volpe, G. (2003). Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59 (1-2), 213–225.

Camurri, A., Lagerlof, I., & Volpe, G. (2003). Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *Int. Journal of Human-Computer Studies*, vol. 59, no. 1-2, 213–225.

Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., & Volpe, G. (2004). Multimodal analysis of expressive gesture in music and dance performances. *Gesture-based Communication in HCI*, 20–39.

Castellano, G., Villalba, S., & Camurri, A. (2007). Recognizing human emotions from body movement and gesture dynamics. *In ACII’07: Proceedings of the Second*

International Conference on Affective Computing and Intelligent Interaction, 71–82.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27: 1-27:27, 2011.

Chang, K.Y., Liu, T.L., & Lai, S.H. (2009). Learning partially-observed hidden conditional random fields for facial expression recognition. *Computer Vision and Pattern Recognition, CVPR 2009. IEEE Conference on*, 533–540.

Chang, Y., Hu, C., & Turk, M. (2004). Probabilistic Expression Analysis on Manifolds. *In Proc. IEEE International' 1 Conf. Computing. Vis. Pattern Recognition*.

Chavan, U.B., & Kulkarni D. B. (2013). Facial Expression Recognition - Review. *International Journal of Latest Trends in Engineer-ing and Technology* (IJLTET), Vol. 3, No. 1, pp. 237–243.

Clarke, T.J., Bradshaw, M.F., Field, D.T., Hampson, S.E., & Rose, D. (2003). The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception*, 34, 1171–1180.

Cohn, J., Kreuz, T., Yang, Y., Nguyen, M., Padilla, M., Zhou, F., & Fernando, D. (2009). Detecting depression from facial actions and vocal prosody. *In proceeding: International Conference on Affective Computing and Intelligent Interaction (ACII2009).*

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

Cowie, R. & Sawey, M. (2011). GTrace - General trace program from Queen's, Belfast, *http://www.dfki.de/~schroed/feeltrace/.*

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.

Darwin, C. (1872). The Expression of Emotions in Man and Animals. *Murray, London, reprinted by University of Chicago Press*, 1965.

David, A.F. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Davis, RA., Charlton, A., Oehlschlager, S., & Wilson, J. (2006). Novel feature selection method for genetic programming using metabolomics 1H NMR data. *Chemo metrics and Intelligent Laboratory Systems*, 81 (1), 50–59.

De Gelder, B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society*, 364 (3), 3475–3484.

De Gelder, B., Snyder, J., Greve, D., Gerard, G., & Hadjikhani, N., (2003). Fear fosters flight: A mechanism for fear contagion when perceiving emotion expressed by a whole body, *Proc. of the National Academy of Science*, 101 (47), 16701–16706.

De Silva, R. & Bianchi-Berthouze, N. (2004). Modeling human affective postures: An information theoretic characterization of posture features. *Journal of Computational Agents and Virtual Worlds*, 15 (3-4), 269–276.

DeGroot, M. H., & Schervish, M. J. (2011). Probability and Statistics (4th edition). Published by Pearson.

Dornaika, F., Lazkano, E., & Sierra, B. (2011). Improving dynamic facial expression recognition with feature subset selection. *Pattern Recognition Letts*, 32 (5), 740–748.

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 Affective Computing and Intelligent Interaction: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, Lisbon, Portugal, 488–500.

Ekman, P., & Friesen, W.V. (1967). Head and body cues in the judgment of emotion: A reformulation. *Perceptual and Motor Skills*, vol. 24, 711–724.

Ekman, P., & Friesen, W.V. (1967). Pictures of Facial Affect. *Consulting*, Psychologists Press.

Ekman, P., & Friesen, W.V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17 (2), 124–129.

Ekman, P., & Friesen, W.V. (1983). Emfacs-7: Emotional Facial Action Coding System. *University of California at San Francisco*.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). Facial Action Coding System, the Manual. *Published by Research Nexus division of Network Information Research Corporation*, USA.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). Facial Action Coding System Investigator’s Guide. *Consulting Psychologist Press*, Palo Alto, CA.

Farid, D. M., & Rahman, C. M. (2013). Assigning weights to training instances increases classification accuracy. *International Journal of Data Mining & Knowledge Management Process*, issue 3, 13–25.

Farid, D., Zhang, L., Hossain, A.M., Rahman, C.M., Strachan, R., Sexton, G., & Dahal, K. (2013). An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams. *Expert Systems with Applications*, 40 (15). 5895–5906.

Fellous, J. & Arbib, M.A. (2005). *Who needs emotions? The brain meets the robot*. Oxford University Press.

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

G'Mussel, A.S., & Hewig, J. (2013). The value of a smile: Facial expression affects ultimatum-game responses. *Judgment and Decision Making*, 8 (3), 381–385.

Garc´ıa-Pedrajas, N., Herv ́as-Mart´ınez, C., & Ortiz-Boyer, D. (2005). Cooperative Coevolution of Artificial Neural Network Ensembles for Pattern Classification. *IEEE Transactions on Evolutionary Computation*, 9 (3), 271–302.

Giese, M.A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Neuroscience*, vol. 4, 179–191.

Goldberg, D.E. (1989). Genetic algorithms in search, optimization and machine learning. *Addison-Wesley Longman Publishing Co., Inc. Boston*, MA, USA.

Gong, B., Wang, Y., Liu, J., & Tang, X. (2009). Automatic Facial Expression Recognition on A Single 3D Face by Exploring Shape Deformation. *ACM Multimedia*, 569–572.

Gu, W., Xiang, C., Venkatesh, Y.V., Huang, D., & Lin, H. (2012). Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, 45, pp. 80–91.

Gunes, H. & Pantic, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *IEEE Trans on Systems, Man, and Cybernetics, Part B*, 39 (1), 64–84.

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

Gunes, H., Piccardi, M., & Pantic, M. (2008). From the lab to the real world: Affect recognition using multiple cues and modalities. *In Jimmy Or (Ed.), Affective computing, focus on emotion expression, synthesis and recognition* (pp. 185–218).

Vienna, Austria: I-Tech Education and Publishing.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, 16 (9), 458–466.

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced computer vision with Microsoft Kinect sensor: A review. *Cybernetics, IEEE Transactions on*, 43(5), 1318–1334.

Harrigan, J., Rosenthal, R., & Scherer, K. (2005). The new handbook of Methods in Nonverbal Behavior Research, Oxford University Press.

Hecht-Nielsen, R. (1989). Theory of the Backpropagation neural network. *Neural Networks, IJCNN, International Joint Conference on,* San Diego, CA, USA.

Hsu, C., Chang, C., & Lin, C. (2010). A practical guide to support vector classification. Department of Computer Science National, Taiwan University.

Huang, C.L. & Wang, C.J. (2006). A GA-based feature selection and parameters optimization for support vector machine. *Expert Systems with Applications*, 31 (2), 231–240.

Izard, C.E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological bulletin*, 115 (2), 288–299.

Izard, C.E., Ackerman, P.B., Schoff, K.M., & Fine, S.E. (2000). Self-organization of discrete emotions, emotion patterns, and emotion cognition relations. *Emotion, Development, and Self-Organization: Dynamic Systems Approaches to Emotional Development*, D.L. Marc and I. Granic (Eds). Cambridge University Press. 15–36.

Jeatrakul, P. & Wong, K.W. (2009). Comparing the performance of different neural networks for binary classification problems. Natural Language Processing, *SNLP '09. Eighth International Symposium on*, 111–115.

Jeong, D.H., Ziemkiewicz, C., Ribarsky, W., and Chang, R. (2009). Understanding Principal Component Analysis Using a Visual Analytics Tool. *Charlotte Visualization Center*, UNC Charlotte.

Kaltwang, S., Rudovic, O., & Pantic, M. (2012). Continuous pain intensity estimation from facial expressions. In Advances in Visual Computing. Lecture Notes in Computer Science, vol. 7432. Heidelberg: Springer, 368–377.

Kamisato, S., Odo, S., Ishikawa, Y., & Hoshino, K. (2004). Extraction of motion characteristics corresponding to sensitivity information using dance movement. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 8 (2), 167–178.

Kanluan, I., Grimm, M., & Kroschel, K. (2008). Audio-visual emotion recognition using an emotion recognition space concept. *Proc. of European Signal Processing Conference.*

Kapoor, A., Burleson, W., & Picard, R.W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65 (8), 724–736.

Kappas, A. (2010). Smile When You Read This, whether you like it or not: Conceptual Challenges to Affect Detection. *IEEE Transactions on Affective Computing*, 1 (1), 38–42.

Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., & Driessen, P.F. (2005). Gesture-based affective computing on motion-capture data. *In Proceeding of First International Conference on Affective Computing and Intelligent Interaction*, 1–7.

Karg, M., Kuhnlenz, K., & Buss, M. (2010). Recognition of affect based on gait patterns. *IEEE Trans on Systems, Man, and Cybernetics, Part B*, 40 (4), 1050–1061.

Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., & Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. *In: Lecture notes in artificial intelligence*, vol. 4451, Springer, Berlin, 91–112.

Kim, J. (2007). Robust Speech Recognition and Understanding. *Bimodal Emotion Recognition using Speech and Physiological Change*s, I-Tech Education and Publishing, 265–280.

Kleinsmith, A. & Bianchi-Berthouze, N. (2007). Recognizing Affective Dimensions from Body Posture. Proc *Int. Conf. of Affective Computing and Intelligent Interaction*, 48–58.

Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *Affective Computing, IEEE Transactions on*, 4 (1), 15–33.

Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). Automatic Recognition of Non-Acted Affective Postures. *IEEE Trans. on Systems, Man, and Cybernetics Part B*, 41 (4), 1027–1038.

Kleinsmith, A., Fushimi, T., & Bianchi-Berthouze, N. (2005). An incremental and interactive affective posture recognition system. *In Proceedings: International Workshop on Adapting the Interaction Style to Affective Factors*.

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), 1940–1954.

Kotsia, I., Zafeiriou, S., & Pitas, I. (2008). Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition*, 41 (3), 822–851.

Kraipeerapun, P. (2008). Neural network classification based on quantification of uncertainty. *Ph.D. thesis,* Murdoch University.

Lange, J. & Lappe, M. (2007). The role of spatial and temporal information in biological motion perception. *Advances in Cognitive Psychology*, vol. 3, no. 4, 419–428.

Li, Y., Chen, J., Zhao, Y., & Ji, Q. (2013). Data-free Prior Model for Facial Action Unit Recognition. *Affective Computing, IEEE Trans-actions on*, 4 (2), 127–141.

Lobo, FG., Goldberg, DE., & Pelkian, M. (2000). Time complexity of genetic algorithms on exponentially scaled problems. IlliGAL Report No. 2000015. Illinois Genetic Algorithm Laboratory, University of Illinois.

Lucey, P., Cohn, J., Lucey, S., Matthews, I., Sridharan, S., & Prkachin, K. (2009). Automatically Detecting Pain Using Facial Actions. *In Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 1–8.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, *In Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis*, San Francisco, USA.

Lucey, S., Matthews, I., Hu, C., Cohn, J., & Ambadar, Z. (2006). AAM derived face representations for robust facial action recognition. *In IEEE International Conference on Automatic Face and Gesture Recognition* (FGR2006).

Mahoor, M. H., Zhou, M., Veon, S., & Cohn, J. (2011). Facial action unit recognition with sparse representation. *In Proc. IEEE Automatic Face & Gesture Recognition and Workshops* (FG 2011).

Masud, M.M., Gao, J., Khan, L., Han, J., & Thuraisingham, B. (2011). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23, 859–874.

Mendes-Moreira, J.A., Soares, C., Jorge, A.M., & Sousa, J.F.D. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys*, 45(1), 1–40.

Meng, H., Kleinsmith, A., & Bianchi-Berthouze, N. (2011). Multi-score Learning for

Affect Recognition: The Case of Body Postures. *Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, Volume 6974, 225–234.

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

Microsoft Corporation. (2013). Kinect for windows SDK programming guide, version 1.8.

Mitchell, T., & Hill, M. (1997). Machine Learning. Publisher McGraw-Hill, Inc., New York, USA.

Montepare, J., Koff, E., Zaitchik, D., & Albert, M. (1999). The use of body movements and gestures as cues to emotions in younger and older adults. *Journal of Nonverbal Behavior*, 23 (2), 133–152.

Mpiperis, L. (2008). 3D facial expression recognition using swarm intelligence. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Las Vegas, Nevada, USA.

Nicolaou, M., Gunes, H., & Pantic, M. (2010). Audio-visual classification and fusion of spontaneous affective data in likelihood space. in *Proc. of IEEE Int. Conf. on Pattern Recognition,* 3695–3699.

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

Oh, I.S., Lee, J.S., & Moon, B.R. (2004). Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (11), 1424–1437.

Owusu, E., Zhan, Y., & Mao, Q.R. (2014). A neural-AdaBoost based facial expression recognition system. *Expert Systems with Applications*, 41 (7), 3383–3390.

Pandzic, I., & Forchheimer, R. (2002). MPEG-4 Facial Animation: the Standard, Implementation and Applications. Wiley.

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

Papadimitriou, C. H. (1994). *Computational Complexity*. Addison-Wesley.

Park, H., Park, J., Kim, U., & Woo, W. (2004). Emotion recognition from dance image sequences using contour approximation. *LNCS: Proc. Int'l Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, 547–555.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 27, 1226–1238.

Pinto, S.C.D., Mena-Chalco, J.P., Lopes, F.M., Velho, L., & Cesar, R.M. (2011). 3D Facial Expression Analysis by Using 2D and 3D Wavelet Transforms. *18th IEEE International Conference on Image Processing (ICIP)*, 1281–1284.

Pollick, F.E., Paterson, H.M., Bruderlin, A., & Sanford, A.J. (2001). Perceiving affect from arm movement. *Cognition*, 82 (2), B51–B61.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17 (3), 715–734.

Rao, K.S., Saroj, V.K., Maity, S., & Koolagudi, S.G. (2011). Recognition of emotions from video using neural network models. *Expert Systems with Applications*, 38 (10), 13181–13185.

Roether, C., Omlin, L., Christensen, A., & Giese M.A. (2009). Critical features for the perception of emotion from gait. *Journal of Vision*, 8 (6), 1–32.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, Volume 33, Issue 1-2, 1–39.

Ryan, A., Cohn, J., Lucey, S., Saragih, J., Lucey, P., & Rossi, A. (2009). Automated Facial Expression Recognition System. *In Proceedings of the International Carnahan Conference on Security Technology*, 172–177.

Rylander, B., & Foster, J. (2000). GA-hard problems. *Proc. Genetic and Evolutionary Computation Conference*, 2000.

Rylander, B., & Foster, J. (2001). Computational Complexity and Genetic Algorithms. *Proceedings of the World Science and Engineering Society's Conference on Soft Computing, Advances in Fuzzy Systems and Evolutionary Computation*, 248–253.

Salahshoor, S. & Faez, K. (2012). 3D Face Recognition Using an Expression Insensitive Dynamic Mask. *Image and Signal Processing, lecture Notes in Computer Science*, Volume 7340, 253–260.

Sandbach, G., Zafeiriou, S., Pantic, M., & Rueckert, D. (2012). Recognition of 3D facial expression dynamics. *Image Vision Computing*, issue 30, 762–773.

Sandbach, G., Zafeiriou, S., Pantic, M., & Rueckert, D. Recognition of 3D facial expression dynamics. *Image Vision Compute*, 30 (10), 762–73.

Savran, A., Alyuz, N., Dibeklioglu, H., Celiktutan, O., Gokberk, B., Sankur, B., & Akarun, L. (2008). Bosphorus database for 3D face analysis. In: Proc. *First COST 2101 Workshop on Biometrics and Identity Management*, Denmark, 47–56.

Savran, A., Sankur, B., & Bilge, M.T. (2012). Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units. *Pattern Recognition*, 45 (2), 767–782.

Savran, A., Sankur, B., & Bilge, M.T. (2012). Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30 (10), 774–784.

Savva, N., Scarinzi, A., & Bianchi-Berthouze, N. (2012). Continuous Recognition of Player's Affective Body Expression as Dynamic Quality of Aesthetic Experience. *IEEE Transactions on Computational Intelligence and AI in Games*, 4 (3), 199–212.

Scherer, K.R. & Wallbott, H.G. (1990). Ausdruck von Emotionen. *Published by Hogrefe*. Chapter 6, 345–422.

Sha, T., Song, M., Bu, J., Chen, C., & Tao, D. (2011). Feature level analysis for 3D facial expression recognition. *Neurocomputing*, 74 (12-13), 2135–2141.

Shan, C., Gong, S., & McOwan, P.W. (2009). Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study. *Image and Vision Computing*, vol. 27, 803–816.

Sikora, R., & Piramuthu, S. (2007). Framework for efficient feature selection in genetic algorithm based data mining. *European Journal of Operational Research*, 180 (2), 723–737.

Soltani, A.R., Tawfik, H., Goulermas, J.Y., Fernando, T. (2002). Path planning in construction sites: performance evaluation of the Dijkstra, A*, and GA search algorithms. *Advanced Engineering Informatics*, 16 (2002), 291–303.

Sorci, M., & Thiran, J.Ph. (2010). Modeling human perception of static facial expressions. *Image and Vision Computing*, 28 (5). 790–806.

Soyel, H. & Demirel, H. (2009). Optimal feature selection for 3D facial expression recognition with geometrically localized facial features. *In: Proc. Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*, Famagusta, Cyprus.

Soyel, H. & Demirel, H. (2010). Optimal feature selection for 3D facial expression recognition using coarse-to-fine classification. *Turkish Journal of Electrical Engineering and Computer Sciences*, 18 (6), 1031–1040.

Soyel, H., & Demirel, H. (2007). Facial Expression Recognition Using 3D Facial Feature Distances. *Lecture Notes in Computer Science Volume 4633, ICIAR07*, 831–838.

Srivastava, R., & Roy, S. (2009). 3D facial expression recognition using residues. *In TENCON, 2009 IEEE Region 10 Conference*, 1–5.

Stein, B. and Meredith, M.A. (1993). *The Merging of Senses*. USA, MIT Press.

Suwa, M., Sugie, N., & Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. *In: International Joint Conference on Pattern Recognition*, 408–410.

Swets D.L., & Weng, J. (1996). Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Transactions on Pattern Analysis and Ma-chine Intelligence,* 18 (8), pp. 831–836.

Tang, H., & Huang, T. S. (2008). 3D facial expression recognition based on properties of line segments connecting facial feature points. *In Proc. 8th IEEE International Conf. Automatic Face Gesture Recognition*, 1–6.

Tang, H., & Huang, T. S. (2008). 3D facial expression recognition based on automatically selected features. *Computer Vision and Pat-tern Recognition Workshops, IEEE Computer Society Conference on*, 1–8.

Tekgüç U., Soyel, H., & Demirel, H. (2009). Feature selection for person independent 3D facial expression recognition using NSGA-II. *in: Proc. 24rd International Symposium on Computer and Information Sciences*, Guzelyurt, Turkey, 35–38.

Tian, Y., Kanade, T., & Cohn, J.F. (2005). Facial Expression Analysis. *Handbook of Face Recognition*, Springer New York.

Tong, Y., Chen, J., & Ji, Q. (2010). A Unified Probabilistic Framework for Spontaneous Facial Action Modeling and Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 32 (2), 258–273.

Tong, Y., Liao, E., & Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans-actions on Pattern Analysis and Machine Intelligence,* 29 (10), 1683–1699.

Torralba, A., & Efros, A. (2011). Unbiased Look at Dataset Bias. *IEEE Conference on Computer Vision and Pattern Recognition*.

Tsalakanidou, F., & Malassiotis, S. (2010). Real-time 2D+3d Facial Action and Expression Recognition. *Pattern Recognition*, 43 (5), 1763–1775.

Ujir, H. (2013). 3D facial expression classification using a statistical model of surface normals and a modular approach. *Ph.D. thesis*, University of Birmingham.

Valstar, M., & Pantic, M. (2006). Biologically vs. logic inspired encoding of facial actions and emotions. *In Proc. of IEEE Intl. Conf. on Multimedia and Expo* (ICME), 325–328.

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

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.

Van den Stock, J., Righart, R., & de Gelder, B. (2007). Body expressions influence recognition of emotions in the face and voice, *Emotion*, 7 (3), 487–494.

Vania, L.M., Lemay, M., Bienfang, D.C., Choi, A.Y., & Nakayama. K. (1990). Intact biological motion and structure from motion perception in a patient with impaired motion mechanisms: A case study. *Visual Neuroscience*, vol. 5, 353–369.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer Verlag, USA.

Vapnik, V. N. (2001). The nature of statistical learning theory, 2nd edition. Springer, New York, USA.

Velusamy, S., Kannan, H., Anand, B., Navathe, B., & Sharma, A. (2011). A Method to Infer Emotions From Facial Action Units. *In: IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP (2011).

Vukadinovic, D., & Pantic, M. (2005). Fully automatic facial feature point detection using Gabor feature based boosted features. In Proc. IEEE Int. Conf. Syst., Man, and Cybernetics, 1692–1698.

Wallbott, H.G. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28 (6), 879–896.

Wang, J., Yin, L., Wei, X., & Sun, Y. (2006). 3D facial expression recognition based on primitive surface feature distribution. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Washington, DC, USA.

Wang, T. H., & Lien, J. (2009). Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation. *Pattern Recognition*, vol. 42, 962–977.

Wang, X., Yang, J., Teng, X., Xia, W., & Jensen, R. (2007). Feature selection based on rough sets and particle swarm optimization. *Pattern Recognition Letters*, 28 (4), 459–471.

Webb, J., & Ashley, J. (2012). Beginning Kinect programming with the Microsoft Kinect SDK. Published by Apress, USA.

Whitehill, J., & Omlin, C.W. (2006). Haar Features for FACS AU Recognition. *Automatic Face and Gesture Recognition, FGR 2006, 7th International Conference on*, 217–222.

Whitehill, J., Tingfan, W., Fasel, I., Frank, M., Movellan. J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). *IEEE Conference on, Automatic Face & Gesture Recognition and Workshops* (FG 2011), 298–305.

Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. *In Proc. of 9th Inter, speech Conf.*, 597–600.

Zavaschi, T., Oliveira, L.S., Souza Jr, A.B., and Koerich, A. (2013). Fusion of feature sets and classifiers for facial expression recognition. *Expert System with Applications*, 40 (2), 646–655.

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

Zhang, L. (2011). Facial Expression Recognition Using Facial Movement Features. *Affective Computing, IEEE Transactions on*, 2 (4), 219–230.

Zhang, L., Jiang, M., Farid, D., & Hossain, M.A. (2013). Intelligent facial emotion recognition and semantic based topic detection for a humanoid robot. *Expert Systems with Applications*, 40 (13), 5160–5168.