Kent Academic Repository

Improving Multi-view Facial Expression Recognition

in Unconstrained Environments

A Thesis Submitted to the University

of Kent

For the Degree of Doctor of Philosophy

in Electronic Engineering

By

Xuejian Wang

May 2016

Abstract

Facial expression and emotion-related research has been a longstanding activity in psychology while computerized/automatic facial expression recognition of emotion is a relative recent and still emerging but active research area. Although many automatic computer systems have been proposed to address facial expression recognition problems, the majority of them fail to cope with the requirements of many practical application scenarios arising from either environmental factors or unexpected behavioural bias introduced by the users, such as illumination conditions and large head pose variation to the camera. In this thesis, two of the most influential and common issues raised in practical application scenarios when applying automatic facial expression recognition system are comprehensively explored and investigated. Through a series of experiments carried out under a proposed texture-based system framework for multi-view facial expression recognition, several novel texture feature representations are introduced for implementing multi-view facial expression recognition systems in practical environments, for which the state-of-the-art performance is achieved. In addition, a variety of novel categorization schemes for the configurations of an automatic multi-view facial expression recognition system is presented to address the impractical discrete categorization of facial expression of emotions in real-world scenarios. A significant improvement is observed when using the proposed categorizations in the proposed system framework using a novel implementation of the block based local ternary pattern approach.

Acknowledgement

I would like to express my sincere and heartfelt gratitude to my supervisor, Prof. Mike Fairhurst, for his tremendous support during my doctoral study. Thanks him for leading me out of ambiguity and confusion when I was ever in doubt of something, for helping me get through the toughest period with his profound expertise as a researcher, and for being a magnificent lifetime mentor. Without his guidance and encouragement, I would have not been able to complete my Ph. D. study.

Also, I would like to acknowledge all my friends and colleagues in Digital Processing Group, at the School of Engineering and Digital Art, who have offered their help along the way of my Ph. D. studying.

Finally, but not the least I would like to dedicate my deepest appreciation and love to all my family. Thank you all for funding and supporting me in completing this doctoral study.

Table of Contents

List of Figures

List of Tables

List of acronyms

FACS	Facial Action Coding System
EMFACS	Emotional Facial Action Coding System
SURF	Speeded-Up Robust Feature
AAM	Active Appearance Model
ASM	Active Shape Model
PDM	Point Distribution Model
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
GLOH	Gradient Location and Orientation Histogram
HOG	Histogram of Oriented Gradient
DCT	Discrete Cosine Transform
LBP	Local Binary Pattern
ELTP	Elongated Ternary Pattern
ELBP	Elongated Binary Pattern
LOD	Level of Difference
CK+	Extended Cohn-Kanade Database
JAFFE	Japanese Female Facial Expression Database
BU-3DFE	Binghamton University 3D Facial Expression Database
SIFT	Scale Invariant Feature Transform

∏-SIFT	Affine Transform-based SIFT
KLD	Kullback Leibler Divergence
D-SIFT	Dense Scale Invariant Feature Transform
GMM	Gaussian Mixture Model
SSVQ	Supervised Soft Vector Quantization
SPM	Spatial Pyramid
LTP	Local Ternary Pattern
LTP^{ms}	Multi-scale Local Ternary Pattern
BBLTP	Block Based Local Ternary Pattern
BBLTP^{ms}	Block Based Multi-scale Local Ternary Pattern
DWT	Discrete Wavelet Transform
EHMM	Ergodic Hidden Markov Model
LGBP	Local Gabor Binary Pattern
SVM	Support Vector Machine
GLCM	Gray Level Co-occurrence Matrix
BBLBP	Block Based Local Binary Pattern
BBLBP^{ms}	Block Based Multi-scale Local Binary Pattern

Chapter 1

Introduction

This chapter describes in detail the general research background of this thesis and presents the key theories and developments in researches into facial expression and emotion which it addresses, which have a profound implication and significant influence on the development of automatic facial expressions recognition technology. Also, this chapter reviews the state-of-the-art in automatic recognition systems of facial expressions in the literature (although further, and more detailed topic reviews will also be included in relevant chapters) and further explains the applications of such a technology in practice along with its limitations. Finally, this chapter summarize the objectives, key contributions, and organization of this thesis.

Section 1.1 briefly introduces the background of this research. Section 1.2 provides an overall background of the development of research studies in facial expression and emotion and key theories and techniques which influence the development of automatic facial expression recognition technology. Section 1.3 describes the general structure and key components of an automatic facial expression recognition system. Also, a review of the state-of-the-art in facial expression recognition systems is included in this section. Section 1.4 describes the current and potential applications of automatic facial expression recognition technology. Section 1.5 summarizes a range of desirable features of an ideal facial expression recognition system and the corresponding limitations that are addressed by these features. Section 1.6 explains the objectives and key contributions of this thesis. Section 1.7 presents the organization of this thesis.

1.1 Introduction

Since the initial important studies of facial expression by Bell [1], Duchenne [2], and Darwin [3], facial expression-related research has continued for over a century. In the 1960s, Izard [4] and Tomkins [5] postulated the discrete theory of emotions. In 1978, Ekman and Friesen introduced the Facial Action Coding System (*FACS*) [6], and the emotional facial action coding system (*EMFACS*) [7] in 1983 which attempted to describe facial expression by analysing various facial actions, such as eyebrow raising, and mouth opening. In 1980, Russell [8] postulated a circumplex model of emotion, which argued that discrete states occur in emotions and systematically described two prominent dimensions in the affective space. In 1993, Ekman [9], [10], introduced six prototypical facial expression families, which are considered universal across all cultures and ethnic groups. With enormous effort, and contributions from emotional theorists and experimental psychologists, the science underlying the facial expressions and emotions of humans is understood better than it ever has been, although a consensus has not yet been reached on the fundamental theories of facial expression and emotion.

More recently, research on automatic recognition of facial expression has become a very active research area within the computer vision and pattern recognition community, largely because of its wide range of potentially important applications in many areas, including affective human computer interaction, artificial intelligence and the robotics industry, patient care monitoring, diagnosing psychiatric diseases, and so on. However, some practical issues still greatly affect the application of facial expression recognition technology in practical scenarios, including illumination variations, head pose variations, accessory inclusions, discrete categorization of facial expression, facial region occlusions, and so on [11]. Without considering and resolving these issues, applications of automatic recognition of facial expression technology in a practical environment are often impractical. In the study reported in this thesis, a cohesive and comprehensive exploration of automatic facial expression recognition is presented to resolve some of the aforementioned issues arising in

practical application scenarios. Besides, several approaches are presented to resolve the encountered impediments, reduce the general difficulty of application of such technology in practice, and introduce some novel, continuous, and generic categorizations of facial expressions for practical applications.

The rest of this chapter will thoroughly review facial expression recognition technology, the primary challenges, and issues faced in the recent development of facial expression recognition systems. Then, the main objectives and key contributions of this study are discussed in detail, followed by a brief description of the organization and content of this thesis.

1.2 A consideration of some applications

Practically, automatic facial expression recognition technology has a range of potential or immediately useful practical applications in various areas, some of which can be listed as follows:

- Replacing manual reading and coding of facial actions: one of the immediate applications of automatic facial expression recognition technology is to replace the manual reading and coding of facial expressions using the *FACS* in a range of psychological studies, which will significantly reduce the time required for analyzing facial actions and thus benefit the research community undertaking studies in psychology.

- Diagnosing psychiatric diseases: automatic facial expression recognition technology can also be used to diagnose certain psychiatric conditions which are related with emotional experience, and monitor psychological states in psychiatric patients [16]. For example, Wang [104] presents a framework for analyzing neuropsychiatric disorders, such as schizophrenia and Asperger's

syndrome.

- Engineering socially affective robots: recent advances in artificial intelligence have made the robotic industry an intensively prosperous and rapidly developing research area. Some interactive robots have been launched to communicate with humans in helping to accomplish a range of tasks [105]. It is rational to enable these robots to read the emotional state of a user, since it has been found that facial expression is not only the most important information been conveyed during a human-to-human conversation, but facial expressions can also even alter the information which is exchanged in words [106]. Without reading a facial expression, a robot is evidently being excluded from the most important information being conveyed from the human face, and therefore fails to establish the fundamental emotional context of a conversation or interaction.

- Designing affective human computer interaction: automatic facial expression recognition technology can also be deployed to improve the design of a human computer interaction and adjust the system interface for affective personalization in order to seek and implement a better user experience [107], [108].

- Improving face recognition systems: an automatic facial expression recognition system can also be used to analyze the facial expression that is projected by a user of a face recognition system, in order to compensate for the errors often introduced by the occurrence of unexpected facial expressions.

- Other applications: an automatic facial expression recognition system can also be seen to have a range of commercial applications. It can be adopted in order to monitor the drivers of vehicles to generate an alarm signal to alert to potential risks, such as drowsiness and eye closure. Additionally, a facial expression recognition system can be utilized to monitor patients who need close care, and signify potential incidents that need to be dealt with or which require a carer to be summoned urgently. Furthermore, a facial expression recognition system may also have applications in marketing research for commercial advertisements [85],

and so on.

1.3 Facial expression of emotion and the *FACS*

There has been a longstanding debate on the perception of facial expressions to convey emotion [12]. With over a century's exploration and study, psychologists have not yet reached a consensus on how facial expression is perceived and what an emotion actually is, and even the terms used in describing emotion are often confused to a large degree [13]. However, research relating to the analysis of facial expression has made significant advances following the introduction of objective measurement tools for facial movements [6], [14], [15]. Among those facial measurement systems, the Facial Action Coding System [6] proposed by Ekman et al. is one of the most influential events in research relating to the automatic recognition of facial expression.

Based on the anatomy of the face, as shown in Figure 1.1, the Facial Action Coding System [6] suggests a set of 44 basic single facial action units and various types of movements of eye and head, and designates each action unit with a specific numeric code for encoding the facial motions read on an expressive face. Additionally, this research also introduced emotion dictionaries (*FACS*/*EMFACS* dictionary and *FACS* Affect Interpretation Database), which is based on their research findings about the categories of facial expressions, to use with the *FACS* to interpret facial expressions categorically [16]. With the support of *FACS*, over 7,000 combinations of facial actions have been observed by researchers [17]. Tables 1.1 and 1.2 provide a list of 46 action units (i.e. consisting of the 44 basic facial actions units and 2 further action units for the eyes) and their numeric codes, and Table 1.3 presents some images of single and combinations of facial action units.
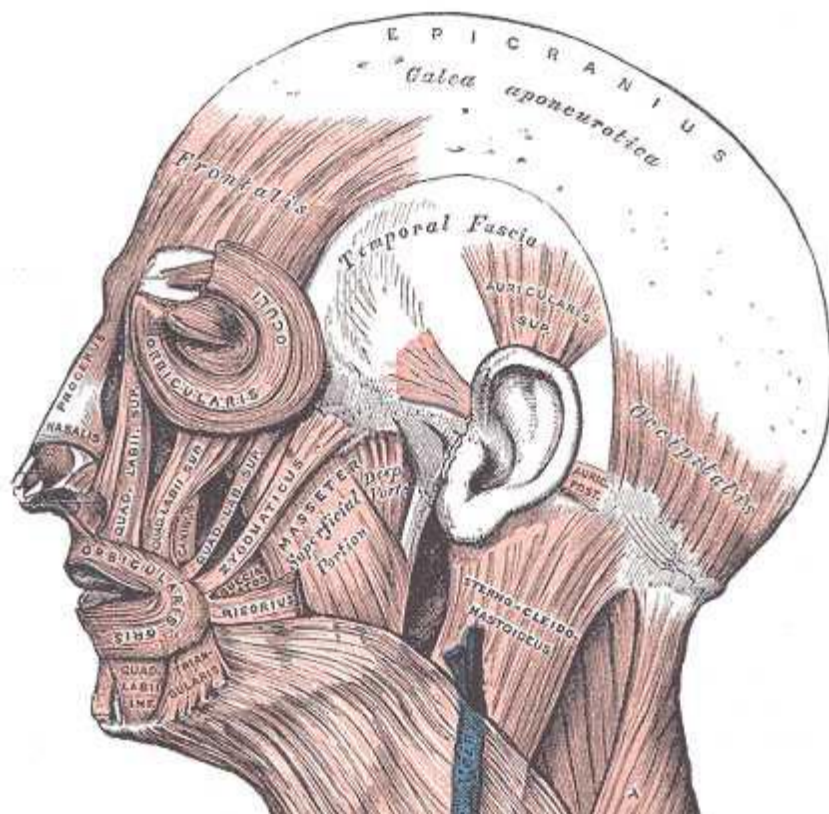
Figure 1.1: An illustration of facial muscles from Gray's Anatomy (taken from [18]).

AU number	Descriptor	Muscular Basis
1.	Inner Brow Raiser	Frontalis, Pars Medialis
2.	Outer Brow Raiser	Frontalis, Pars Lateralis
4.	Brow Lowerer	Depressor Glabellae, Depressor Supercilli; Corrugator
5.	Upper Lid Raiser	Levator Palpebrae Superioris
6.	Cheek Raiser	Orbicularis Oculi, Pars Orbitalis
7.	Lid Tightener	Orbicularis Oculi, Pars Palebralis
9.	Nose Wrinkler	Levator Labii Superioris, Alaeque Nasi
10.	Upper Lip Raiser	Levator Labii Superioris, Caput Infraorbitalis
11.	Nasolabial Fold Deepener	Zygomatic Minor
12.	Lip Corner Puller	Zygomatic Major
13.	Cheek Puffer	Caninus
14.	Dimpler	Buccinator
15.	Lip Corner Depressor	Triangularis
16.	Lower Lip Depressor	Depressor Labii
17.	Chin Raiser	Mentalis
18.	Lip Puckerer	Incisivii Labii Superioris; Incisivii Labii Inferioris
20.	Lip Stretcher	Risorius
22.	Lip Funneler	Orbicularis Oris
23.	Lip Tightener	Orbicularis Oris
24.	Lip Pressor	Orbicularis Oris
25.	Lips Part	Depressor Labii, or Relaxation of Mentalis or Orbicularis Oris
26.	Jaw Drop	Masetter; Temporal and Internal Pterygoid Relaxed
27.	Mouth Stretch	Pterygoids; Digastric
28.	Lip Suck	Orbicularis Oris

Table 1.1: Action units in the Facial Action Coding System (taken from [16]).

AU number	FACS name
8.	Lips Toward Each Other
19.	Tongue Out
21.	Neck Tightener
29.	Jaw Thrust
30.	Jaw Sideways
31.	Jaw Clencher
32.	Lip Bite
33.	Blow
34.	Puff
35.	Cheek Suck
36.	Tongue Bulge
37.	Lip Wipe
38.	Nostril Dilator
39.	Nostril Compressor
43.	Eyes Closure
45.	Blink
46.	Wink

Table 1.2: : Grossly defined action units in *FACS* (Taken from [16]).

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

Table 1.3: Images of facial action units (taken from [19]).

In contrast to the *FACS* which serves as a measurement tool for encoding facial expressions, another notion has also been widely acknowledged and has significantly influenced the research of automatic facial expression recognition. That is the notion that facial expressions can be recognized categorically, and that there are six prototypic families of facial expressions, consisting of *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* [9], [20]. Influenced by this finding, a large proportion of researchers consider the problem of automatic recognition of facial expressions as distinguishing facial expressions among these six prototypic facial expression groupings, although the categorical perception of emotion theory contradicts other postulated theories of emotion [8], [21], which suggest that emotion space is continuous, with some prominent dimensions. In our study, the problem of automatic facial expression recognition is approached from both perspectives. Both categorical and dimensional recognition of facial expression and emotion are explored in this thesis.

1.4 Automatic facial expression recognitions

Following Suwa et al.'s [22] early attempt to recognize facial expressions by analysing regions of interest in an image sequence, the research relating to automatic facial expression recognition has made enormous advances in the past decades. Influenced by two important theories in facial expression, as described in Section 1.2, generally computer-based systems addressing the automatic recognition of facial expressions fall into one of two categories. One category describes systems which attempt to recognize facial expressions as belonging to one of the specified prototypic facial expressions [23], [24], [25]. The other categories applies the component analysis of facial expressions of emotion using the *FACS* [16], [26], [27], [28], [29], [30], which first attempts to recognize facial action units, and then relies on the interpretation dictionary (which is explained in Section 1.2) associated with the *FACS* to classify facial expressions into either emotional or non-emotional categories. Although these two approaches differ significantly in addressing the problem, the fundamental methodology is based on the same assertion that facial expressions of emotion are

visually distinguishable and different emotional states are associated with different facial expressions.

1.4.1 General structure of an automatic facial expression recognition system

In general, an automatic facial expression recognition system consists of three important stages one after another: the data acquisition stage, the feature extraction stage, and the classification stage. Figure 1.2 illustrates the general structure of an automatic facial expression recognition system. At each of the stages of the system, a specific range of tasks is executed, involving different functional modules.

- **Data acquisition stage**: the first stage prepares the acquired data for feature extraction. For a static system (i.e. one which recognizes facial expression from a single frame of the image sequence), a face detection algorithm is applied to locate the face region in the image. However, for a dynamic system (i.e. one which analyzes the temporal dynamics of a facial expression to recognize the specific facial expression), the face is initially located using a face detection algorithm on the first frame of the video sequence, and then the face region or face model is tracked throughout the rest of the video [31]. For an action unit-based system, it is the facial action units or facial regions relating to facial action units that are located and tracked instead of the entire face region. In addition, some issues including face registration, illumination normalization, head pose estimation, and accessory inclusion detection, are normally carried out at this stage, depending on the specific design and requirements of a particular system.
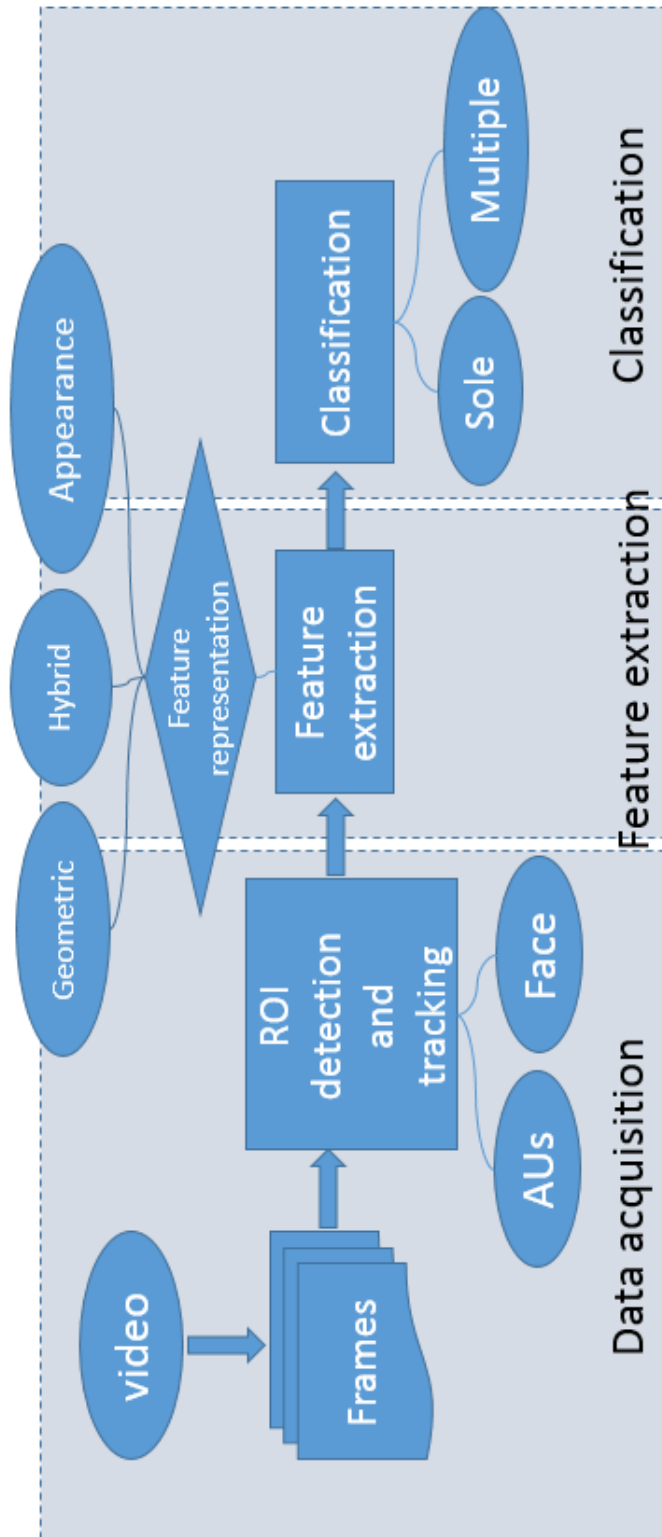
Figure 1.2: General structure of an automatic facial expression recognition system.

● Feature extraction stage: various types of feature extraction algorithm are applied to extract a variety of features to represent facial expression. Depending on the design of the facial expression recognition system, the extracted feature representation can vary from a texture representation to a 3D parameterized facial model. In general, the feature extraction algorithms applied at this stage can be categorized by the type of features that are extracted: geometric based, appearance based, and hybrid feature based extraction algorithms, for example.

● Classification stage: at the last stage of the system, the obtained feature representation of an image or video sequence of facial expression is classified by the system into one of the prototypic facial expressions. Alternatively, for the *FACS-based* automatic facial expression recognition system, a set of presented facial actions are recognized and encoded using *FACS* and then, according to the emotional dictionary, the acquired facial expression is finally interpreted. A single, or a set of multiple of classifiers, can be utilized at this stage depending on the specific design of the system.

1.4.2 Face acquisition

To carry out facial expression recognition, the first step is to locate the face region in the source data (e.g. an image or a video sequence). The Viola and Jones's face detection algorithm [32], [33] is one of the most efficient and fast face detectors which is widely adopted for use in face recognition and facial expression recognition systems. It utilizes Harr-like features, which can be fast computed using the pre-computed integral image, to characterise the face and non-face region and a cascade of boosted weak classifiers to learn the classification rules. The Viola and Jones's face detection algorithm can locate the face image in real time but it also has some disadvantages. For example, training the cascade of classifiers is time-consuming. To further improve the performance of the original Viola and Jones's face detection algorithm, many extensions and improvements have been suggested. Wu et al. [34] and Pham et al. [35] significantly improve the training efficiency and reduce the required training time

of the Viola and Jones's face detection algorithm. Li et al. [36] present another extension, adopting *SURF* features [37] to characterize the face and non-face regions.

In addition, Moghaddam and Pentland [38] devised a visual learning algorithm for face detection based on density estimation. Rowley et al. [39] utilize a set of neural networks to detect a face in small windows, and then re-construct a completed face from the detected small windows that contain face regions. Roth et al. [40] constructed a face detection algorithm that can cope with various head pose and illumination conditions based on a Sparse Network of Winnows (SNoW).
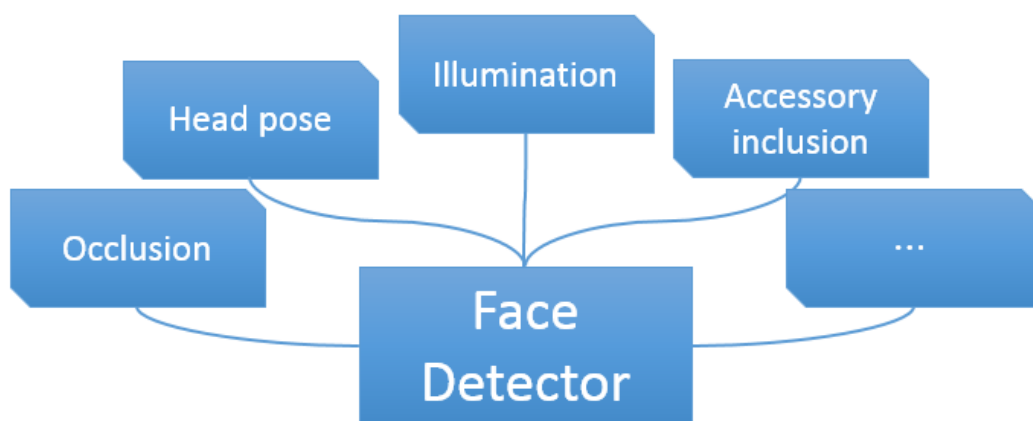
Figure 1.3: Factors that affect the performance of a face detector.

The performance of face detection algorithms in practical scenarios is generally affected by the variations that are caused by illumination variability, large in-plane and out-of-plane rotations [1], occlusions, accessory inclusions, and so on, as illustrated in Figure 1.3. Any of these issues can cause the aforementioned algorithms to fail or detect a false face region. To address these problems, Feraud et al. [41] propose the Constrained Generative Model (CGM) for face detection at frontal view and side

---

[1] In-plane rotation means a rotation that is within 2-dimensional geometric plane; out-of-plane rotation involves 3-dimensional rotation.

views based on neural networks. Schneiderman et al. [42] address the head pose variation issue with a statistical appearance model which can be trained to detect a face at frontal and profile views. Huang et al. [43] propose face detection algorithms which can detect face with in-plane and out-of-plane rotations. Parupati [44] employs a wireless camera network to cope with issues raised by head pose variations. Liao et al. [45] introduce an unconstrained face detection algorithm employing normalized pixel difference (NPD) and a deep quadratic tree that enables efficient and fast learning of classification rules. Forba et al. [46] propose an illumination-invariant local structure feature obtained using the modified Census Transform for face detection under various illumination conditions.

In a dynamic facial expression recognition system, after the face region is successfully detected in the first frame of the video sequence, a tracking algorithm is usually applied to track the established face location or face models to reduce the computation requirement and improve the overall performance of the system [47], [48], [49], [50].

1.4.3 Representations of facial expression

Feature extraction is the second stage in an automatic facial expression recognition system, which examines and extracts the detected regions of interest for generating an informative and discriminative representation for the facial expression to be classified. The literature shows that, generally, there are three types of feature representations that are commonly adopted: geometric features, appearance features, and hybrid features, which possesses properties of both geometric and appearance features.

1.4.3.1 Geometric features

The geometric feature category utilizes the geometric shape information of the facial expression to encode the differences between various facial expressions, for example, the contours of the eyebrows and the inner and outer contours of the mouth. To extract the geometric shape information of a facial expression, a facial landmark localization algorithm is applied to detect fiducial points on a face (e.g. inner corners of eyes, corners of mouth, and nostril) in the detected facial regions. Once the coordinates of these fiducial points are successfully located, a normalization process is usually applied to normalize the size and rotation of the obtained face model. This process is also referred to as the face registration process. It can be achieved by defining a reference face, and then applying a Procrustes analysis [51] of all the located fiducial points or part of the set fiducial points which are less variant across all categories of facial expressions, for example, the inner corners of the eyes and the tip of the nose. However, this registration method is not reliable for dealing with out-of-plane head pose.

A face geometric model is crucial for extracting geometric features, which consists of a set of fiducial points that are defined by an automatic facial expression recognition system. A variety of face geometric models has been proposed in the literature. Raouzaiou et al. [52] present an MPEG-4 compatible framework for analysing and synthesizing of facial expression for human computer interaction, adopting facial animation parameters to model various facial expressions. An illustration of the MPEG-4 facial expression model is presented in Figure 1.4. This sophisticated framework can encode both the previously specified prototypic facial expressions and a variety of facial actions. Ahlberg [53], [54] presents an alternative parameterised face model named Candide-3 which is also compatible with the MPEG-4 standard. Later, Dornaika and Davoine [55] proposed a particle filter-based tracking algorithm for the 3D Candide Model.
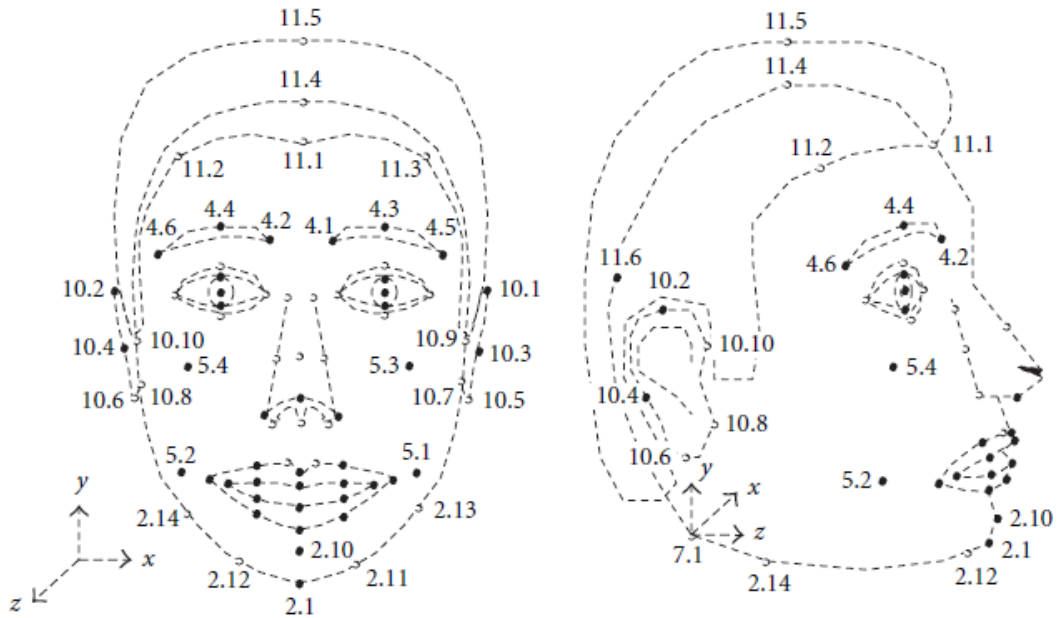
Figure 1.4: A face model of the neutral face using the face model based on MPEG-4 (taken from [52]).

Cootes et al. introduced one of the most dominant and commonly employed algorithms for facial landmark localization, the Active Shape Model (*ASM*) [56], and later the Active Appearance Model (*AAM*) [57], [58]. For the active shape model, they devised a Point Distribution Model (*PDM*) for a geometric shape of a face, which is characterised using a set of landmark points, as shown in Figure 1.5. Then, the Principal Component Analysis is applied on the normalized fiducial points, which are manually labelled, to model the variation of the *PDM* across all training samples of an object and derive a statistical model for the object. The *ASM* algorithm addresses the landmark localization problem as a constrained interactive model-fitting process. At each interaction, after the mean model is initialized on the image, each fiducial point in the face model shifts towards the local edge, and then the new model is constrained using the generalized statistical model. The shift and constrain process continues until

either a specified number of interactions or a convergence is reached. The *AAM* [57] is an extension of the *ASM* algorithm which statistically encodes both the shape and appearance of all training samples to construct the active appearance model for the facial expression image under query. With the deployment of these fundamental techniques, a researcher can generate, locate, and track a face model of their own preference.
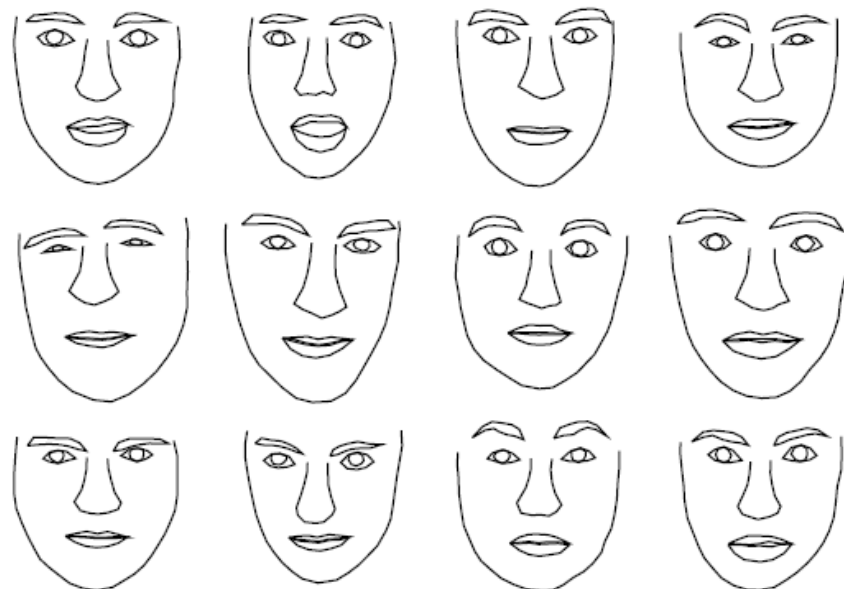
Figure 1.5: Some examples of geometric shape of the face with connected landmark points (taken from [59]).

The original version of the *ASM* and *AAM* algorithms generally find difficulty in handling issues such as slow convergence with poor initialization, poor model fitting under large head pose variations, and failure under partial occlusion of the face. Xiao et al. [60] present a combined 2D and 3D *AAM*, which is constrained by a 3D model, for locating and tracking fiducial points in 3D. Lee and Kim [61] present a tensor-based *AAM* algorithm to overcome the variations raised by subjects, large head pose,

facial expressions, and illumination in localization of fiducial points. Duhn et al. [62] present a three view face tracking algorithm based on the *AAM* which can fit and track the face model in an unconstrained environment, such as surveillance videos, as shown in Figure 1.6. Gross et al. [63] extended the *AAM* algorithm further to handle partial occlusions.
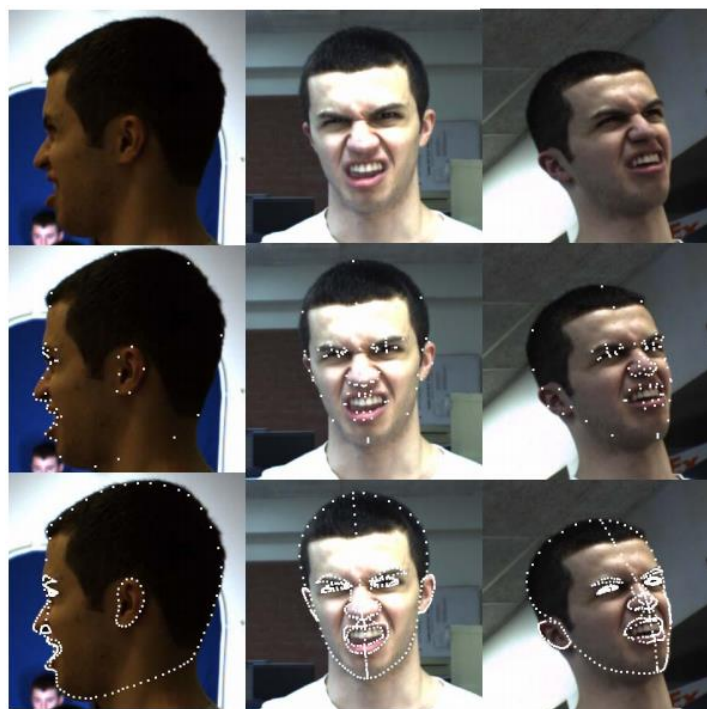
Figure 1.6: Top row: images captured in three different views; middle row: the located face models; bottom row: the expanded face models with interpolated fiducial points (taken from [62]).

Besides the *AAM* and *ASM* based methods, other facial models are also used to encode the displacement of a geometric face model of various facial expressions, including multistate face component models [19], point-base face model of frontal and side view [64], and other facial landmarks based face models [65], [66], [67].

1.4.3.2 Appearance features

Appearance features are generated based on the appearance difference of various types of facial expressions, which utilize a range of texture descriptors to model the appearance of a facial image for a feature representation, such as Gabor Wavelets [68], [69], Principal Component Analysis (PCA) [70], [71], Linear Discriminant Analysis (LDA) [72], Local Binary Pattern (LBP) [73], Speed Up Robust Feature ($SURF$) [37], Gradient Location and Orientation Histogram ($GLOH$) [74], Histogram of Oriented Gradient (HOG) [75], Discrete Cosine Transform (DCT) [76], and so on. In contrast to the case when using geometric features, a minimal feature registration process is applied to normalize the size and in-plane orientation of the face. For example, the centre of the eyes and mouth can be used to align the face image in the query image to the pre-defined reference face [11].

The local binary pattern (LBP) [73] and extensions of LBP [77] operator have been intensively utilized for the analysis of facial expressions in the past decades. Shan et al. [78] adopts the local binary pattern approach for the representation of different facial expressions, and applied template matching and the support vector machine to classify facial expressions. Nanni et al. [79] utilize a combined texture of Elongated Ternary Pattern ($ELTP$) and Elongate Binary Pattern ($ELBP$) for recognizing pain. Liao et al. [80] present another appearance representation using the extended local binary pattern and Tsallis entropy [81] calculated from 16 Gabor filtered images.

Besides the local binary pattern, Deng et al. [82] present a facial expression recognition system based on appearance features generated using Gabor filters for feature generation and PCA for feature selection and reduction. And, Yu et al. [83] present a system using appearance features based on the Weber Local Descriptor and histogram contextualization, and tested on a set of facial expression images generated from a search engine (specifically, Google). Furthermore, Ma et al. [84] utilize the *2D-DCT* algorithm, which was originally introduced specifically for image compression, for implementing a facial expression recognition system.

1.4.3.3 Hybrid features

Both geometric features and appearance features have their own advantages and disadvantages. The geometric feature approach is straightforward to understand and can be computationally efficient in some application scenarios. The registration process can help remove the bias introduced by subject difference and improve the ability to discriminate among facial expressions. However, processes for establishing and tracking the face model generally need to be well designed to cope with challenging issues typically introduced by the application environment and the subject, including changes in illumination, occlusions of face, accessories inclusion, and large variations of head pose. Additionally, the geometric feature approach cannot encode subtle facial expression and some facial actions that are not visually distinguishable in 2D geometric space, such as cheek raising, mouth corner dimpling, and so on [85]. In contrast to the geometric feature approach, the registration methods applied before the extraction of appearance features cannot fully register the face to a reference face properly, and this thus restricts the ability to discriminate among the appearance features. Also, the appearance of a subject is affected by age, make-up applied to the face and ethic group, and thus an appearance feature can be sensitive to variations introduced by subjects' differences noted. Furthermore, general issues that affect the geometric methods also influence the discrimination potential of appearance features, including non-uniform illumination and partial occlusions.

With that being said, it is therefore reasonable and necessary to consider constructing a hybrid feature representation which can embody the inherent strengths of both geometric features and appearance features, so that a superior representation is thereby devised. The resulting hybrid feature approach utilizes both the geometric shape and the appearance of a face to encode the difference among various facial expressions. A hybrid feature representation can be constructed with various types of fusion. For example, it can be an appearance feature extracted at the located landmark points [86] or a more sophisticated representation, which utilizes a geometric face model to normalize the appearance of the face and then applying an algorithm to extract holistic or local appearance features [55], [87], [88], [89].

Kotsia et al. [30] utilize the idea of discriminant non-negative matrix factorization to extract texture information and the Candide grid node information extraction algorithm to extract the shape information of an expressive face. Wang and Yin [90] propose the topographic context-based feature for facial expression recognition, which is extracted at certain regions of the face, as shown in Figure 1.7, after the fiducial points are successfully established.
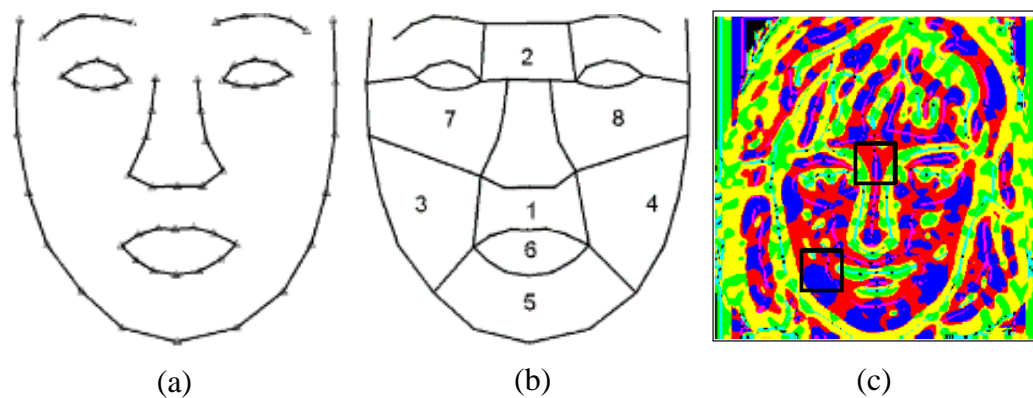
Figure 1.7: (a) Face models of 64 fiducial points; (b) facial expressive regions for extraction of topographic context based features; (c) an example of the terrain map of a face image (taken from [90]).

1.4.4 3-dimensional-based facial expression recognition

As mentioned in Section 1.3.3.3, geometric features cannot encode some subtle facial expressions or facial actions that are reflected in changes of depth - for example, the facial action of cheek raising and mouth corner deepening. Compared with a 2D-based system, a 3D-based system can better interpret the deformation of facial expression in 3D, and with the assistance of a 3D facial acquisition facility, the difficulty encountered when adopting the 2D face model, such as issues introduced by large variation of head pose, can be resolved efficiently. As a result, 3D-based systems

for automatic facial expression recognition have also been an active research area within the scope of general automatic facial expression recognition. A variety of 3D facial expression recognition systems have been proposed, including 3D generic elastic models based [91], 3D deformable model based [92], 3D local shape analysis based [93], histogram of surface differential quantities based [94], local principal curvature analysis based [95] facial expression recognition systems, and others.

1.4.5 Recognition of facial expression in video sequence

In contrast to a static system which recognizes a facial expression in a single frame of a facial image, a dynamic facial expression recognition system can model the dynamic deformation of a facial expression in the dimension of time in order to analyse facial expressions. It has been stressed that the dynamic aspect and timing of facial expressions is important for recognizing spontaneous smiles and can enhance the human perception of subtle facial expression [18], [96], [97]. Therefore, it is worth considering encoding the characteristics of the dynamics of facial expressions for automatic analysis of facial expressions.

Cohen et al. [98] present a dynamic facial expression recognition system based on the Piecewise Bezier Volume Deformation (*PBVD*) tracker, as shown in Figure 1.8, to track the 3D wireframe model of the face in a video in order to model the temporal deformation of various facial expressions, and applied the multi-level Hidden Markov Model for classification of facial expressions.
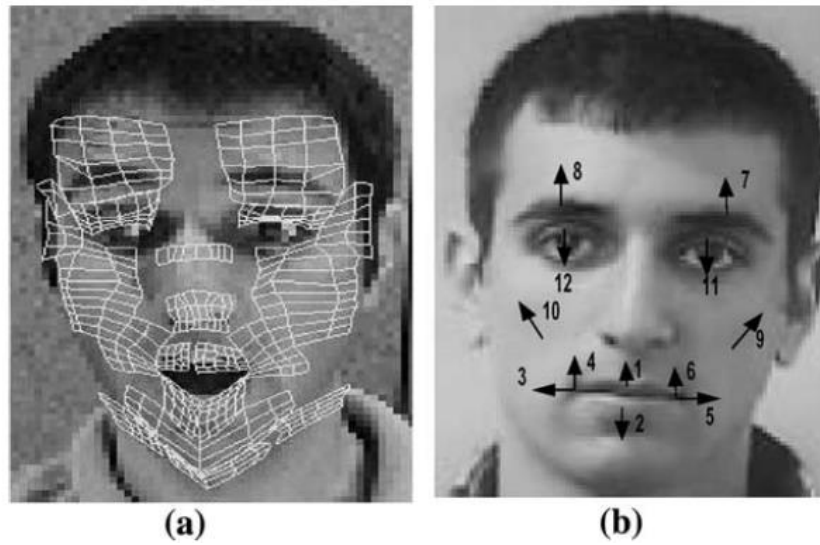
Figure 1.8: (a) The 3D wireframe model used by the *PBVD* tracker; (b) facial deformation encoded for feature representation (taken from [98]).

Li et al. [99] utilize the Candide-3 face model to model the temporal and geometric differences of various facial expressions. Almaev and Valstar [100] introduce the local Gabor binary pattern from the three orthogonal planes (*LGBP-TOP*) for modelling the temporal dynamic of facial expressions and employ the support vector machine for classification of facial expressions. Taini et al. [101] extend the use of the *LGBP-TOP* feature for recognition of facial expressions in near-infrared videos, while Zhao and Pietikainen [24] introduce the volume local binary pattern (*VLBP*) based system to recognize a facial expression, and a comparable performance with *LGBP-TOP* based system is reported. Other dynamic systems, including the appearance-based tracker and the case-based reasoning approach based systems [87], 3D gradient descriptor based systems [102], multistate face component models and Bayesian networks based systems [19], [103], have also been presented for recognizing facial expressions in video sequences.

1.5 Ideal features of an automatic facial expression recognition system

Ideally, the design of an automatic facial expression recognition system should meet a range of requirements that enable and ensure the robustness, efficiency, accuracy, usability, and automaticity of a proposed system as summarized in Table 1.4 (taken from [11]).

Robustness	
Rb1	Deal with subjects of different age, gender, ethnicity
Rb2	Handle lighting changes
Rb3	Handle large head motion
Rb4	Handle occlusion
Rb5	Handle different image resolution
Rb6	Recognize all possible expressions
Rb7	Recognize expressions with different intensity
Rb8	Recognize asymmetrical expressions
Rb9	Recognize spontaneous expressions
Automatic process	
Am1	Automatic face acquisition
Am2	Automatic facial feature extraction
Am3	Automatic expression recognition
Real-time process	
Rt1	Real-time face acquisition
Rt2	Real-time facial feature extraction
Rt3	Real-time expression recognition
Autonomic Process	
An1	Output recognition with confidence
An2	Adaptive to different level outputs based on input images

Table 1.4: Ideal features of an automatic facial expression recognition system (taken from [11]).

An automatic facial expression recognition system should model the facial expressions in a way that is independent of the variation introduced by subjects' incidental differences. For example, age, skin color, and make-up that is worn on the face can influence the performance of an automatic facial expression recognition system which adopts the appearance feature-related algorithm for locating and tracking of fiducial points or generating feature representations.

Similarly, an ideal system should be able to handle the variation of illumination conditions that is common in both practical indoor and outdoor environments, as illumination conditions can significantly affect the quality of the image that is acquired, and thus influence the performance of a system to a large extent, especially in outdoor application scenarios where illumination control is generally very much more difficult to achieve.

In addition, an ideal facial expression recognition system should also be able to cope with various occlusions of the face and accessory inclusion, because these conditions can cause many components of an automatic facial expression recognition system to fail to function normally. For example, an occlusion of the face can cause the face detector to fail to detect a face in an image, the landmark localization algorithm to fail to locate the fiducial points, an appearance-based feature extraction algorithm to extract a false feature representation that contains the occluded face region, and eventually disable the entire facial expression recognition system. Typical examples of partial occlusion and accessory inclusions include the wearing of a scarf, the wearing of glasses, and the wearing of a hat, and so on.

Another desirable feature for an automatic facial expression recognition system is the ability to cope with large in-plane and out-of-plane head pose variations. It is acknowledged that spontaneous facial expression usually occurs along with head motions, especially in outdoor environments [109]. Also, a large variation of head pose alters both the geometric characteristics and appearance of the presented facial

expression, which can eventually significantly influence the performance of a facial expression recognition system. Furthermore, coping with large head pose variation is also an important aspect for improving both the usability and the robustness of a system, particularly for applications in unconstrained environments where effective head pose control measures are not generally applicable. Therefore, it is crucial to take these issues into account when designing and implementing an automatic recognition solution for facial expression analysis.

An ideal system should also be able to recognize all or, at least, a specified range of spontaneous facial expressions that are required by a particular application scenario, even at low intensities. Distinct from a deliberate facial expression which is posed intentionally to show that an emotion is being experienced, a spontaneous facial expression is usually more subtle (i.e. less pronounced) and tends to happen more quickly in a consistent period of time [16]. The majority of automatic facial expression recognition systems which have been proposed are devised based on specifically posed and fully-pronounced (i.e. clearly articulated) facial expressions, an issue which brings about a critical problem. This is that the automatic recognition solution is, in fact, constructed based on biased data, and thus restricts the performance of a laboratory-proven automatic recognition system of facial expression when transferred to practical application environments. With that being said, it is essential to consider the bias introduced by the posed data and to recognize less pronounced facial expression in order to improve the usability and robustness of a devised system in order to achieve higher recognition accuracy when operating subsequently in practical application scenarios.

Another desirable feature of an automatic system is the capability to work in a responsive manner in real time, which suggests that an optimal facial expression recognition system should be able to recognize, at least, one subject's facial expression in a fairly short time, for example, being capable of processing 30 frames of facial expressions within 1 second. Recognizing multiple subjects' facial expressions raises another challenge for designing a real time facial expression recognition system,

which reveals an increase in recognition time, which is linear with respect to the total number of subjects that a system is capable of dealing with. As an example, one typical application scenario of a multiple facial expression recognition system is monitoring psychiatric patients in a communal lounge.

In addition to the factors summarized in Table 1.4, an ideal system should require minimal or no cooperation from the subject and should be capable of recognizing a facial expression without raising concerns on the part of the subject, which means that the system should be able to recognize facial expressions unobtrusively and from a distance.

Also, an automatic facial expression recognition system should be able to interpret facial expressions in a broad and cohesive manner that enables the output emotional label to be appropriately interpreted in a particular context. Currently, the majority of proposed systems have focused on recognizing facial expressions in a discrete way, which categorizes a particular facial expression into one of the prototypic facial expression categories or recognizes the facial expression by coding facial action units in the presented facial expression. This categorization may lead to suboptimal solutions for some practical application when some broader categorization is preferred, and thus a broader and more cohesive categorization can in many cases lead to a better recognition accuracy.

To apply automatic facial expression recognition technology effectively in practice, these aforementioned factors should be carefully considered when devising a practical system for application in a particular scenario. Although features such as these generally result in a better automatic facial expression recognition system, in practice, to design an automatic system that meets all these requirements is infeasible and uneconomic since the requirement of a system is influenced by the particular application scenarios and laid down by application-related system requirements.

Hence, an optimal system in practice is one that fulfils all system requirements with the least resource requirement, taking into account all relevant aspects.

1.6 Objective and key contributions of this research

1.6.1 Aim of this research

Section 1.5 has described a range of limitations and desirable features of an optimal facial expression recognition system, which address the corresponding limitations. In the research to be reported in this thesis, from a broader perspective, the objective of this research is to investigate and resolve the issues relating to facial expression recognition in less constrained or unconstrained environments and make facial expression more interpretable for practical application scenarios.

First, an automatic facial expression recognition system is devised to recognize facial expressions in less constrained or unconstrained environments, under a variety of large head pose variations and a range of expression intensities (this is further explained in Section 2.1.3).

Second, a comprehensive investigation of the application of currently available texture features in multi-view facial expression recognition is carried out to find a novel texture feature or fused texture feature for constructing a better feature representation for multi-view facial expression recognition.

Third, novel categorizations of facial expressions are summarized, which enable the facial expression to be interpreted in a manner that facilitates better and more effective practical application of such a technology.

Fourth, a multi-view facial expression database is designed and constructed initially, which makes the proposed research relating to facial expression recognition in unconstrained environment feasible by making available appropriate data for experimentation and investigation.

1.6.2 Key contributions

Although automatic facial expression recognition has been studied intensively, some questions still have not yet been full addressed by the techniques and systems principally proposed to date. The key contributions of this thesis can therefore be listed as follows:

- A novel automatic recognition system for facial expressions is presented, which is capable of handling facial expressions at 35 different out-of-plane rotations ranging from $+45°$ to $-45°$ pan head rotations and $-30°$ to $+30°$ tilt head rotations. This system also recognizes all six prototypic facial expressions at the least facial expression intensity with significantly better accuracy (i.e. which is measured on 4 different facial expression databases) compared with current state-of-the-art facial expression recognition systems.

- A novel local descriptor is introduced which can be used with other common state-of-the-art texture descriptors, which complement each other in describing the local textures. By fusing the novel texture feature and a range of existing state-of-the-art texture features, we are able to seek a better feature representation for multi-view facial expression recognition systems. It is found that the novel feature when combined with a particular range of texture features significantly improves the overall performance and usability of a system in terms of recognition accuracy compared with a single texture feature based system.

- A novel categorization strategy is also introduced for re-categorizing the range of

possible facial expressions, which interprets the facial expressions in the prominent dimensions of the affective space. This categorization strategy offers a continuous, cohesive, and broader way of interpreting facial expressions with benefits in many practical application scenarios.

1.7 Thesis organization

The thesis is organized as follows:

- In Chapter 2, a self-collected in-house facial expression database that was created for this study along with other public facial expression databases that are deployed in this thesis are comprehensively described. In addition, other techniques, including face detection algorithm and classifier utilized, and toolboxes adopted in this study are presented, to give a complete and accessible summary of the project infrastructure adopted.

- In Chapter 3, a novel universal multi-view facial expression recognition system using local ternary pattern and the extended multi-view local ternary pattern is presented, together with a comprehensive investigation of the influence of the head pose variation and intensity level change of facial expressions on the proposed facial expression recognition system.

- Chapter 4 presents a novel texture descriptor (designated the level of difference (*LOD*) descriptor), which statistically describes the appearance of a local neighborhood, to use with other state-of-the-art texture descriptors. Additionally, a preliminary study of fusion of texture descriptors, including the multi-local ternary pattern, gray scale co-occurrence matrix, local binary pattern, histogram of oriented gradient, and level of difference descriptor, for the design of a universal facial expression recognition system is comprehensively described.

- Chapter 5 reviews the theory of emotion and facial expression and describes the affective models and key dimensions in the emotional space, and then presents a novel strategy for re-categorization of facial expressions for practical application scenarios. A preliminary investigation of the influence of such a novel categorization scheme is conducted based on the proposed multi-view facial expression recognition system that is introduced in Chapter 3.

- Chapter 6 briefly summarizes the content of the thesis, highlights the important observations and findings discovered and reported throughout this study, emphasizes the key contributions of this thesis, and gives suggestions for future research directions.

Chapter 2

Experimental infrastructure

This chapter will describe the public facial expression databases utilized in this study, and present a new multi-view facial expression database in detail, which is designed and collected for devising and evaluating a novel facial expression recognition system. Furthermore, this chapter also explains some essential techniques and toolboxes that are utilized in this study, including the face detection algorithm and the classifier.

Section 2.1 explains the public facial expression database that is employed in this study. Section 2.2 present and thoroughly describes our novel in-house multi-view facial expression database. In Section 2.3, the face detection algorithm and toolbox that is used in this study are described thoroughly. Section 2.4 explains the classifier that is deployed for classification of facial expressions and how the classification accuracy is reported in this study. Section 2.5 briefly summarizes this chapter's content.

2.1 Facial expression databases

This section explains, summarizes, and discusses the facial expression databases that were used in the research study reported, including the Extended Cohn-Kanade Database (*CK+*), the Binghamton University 3D facial expression database, and the Japanese female facial expression database. All of them are available publicly to the academic research community and employed popularly in devising and testing many state-of-the-art facial expression recognition systems. Additionally, a novel facial expression database, our in-house multi-view facial expression database, is designed and collected specially for this study, which is particularly aimed at multi-view facial expression recognition research. The detailed design, data collection setup, and procedures are also presented.

2.1.1 Japanese female facial expression database (*JAFFE*)
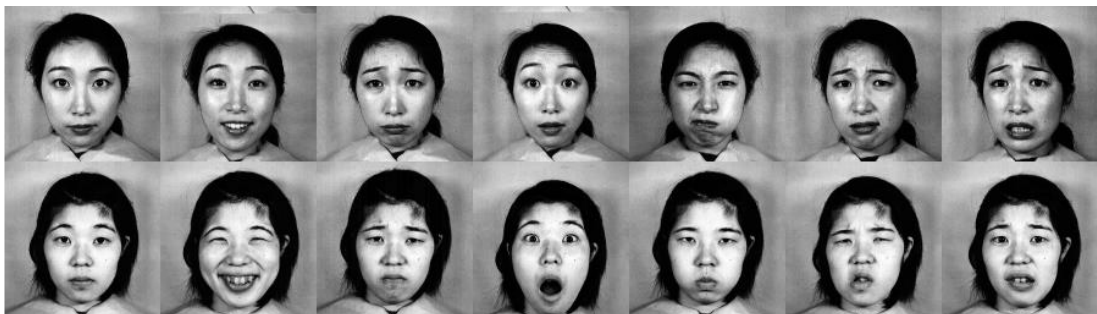
Figure 2.1: Some examples of facial expression images of two subjects in the *JAFFE* database, from left to right demonstrate neutral, happiness, sadness, surprise, anger, disgust, and fear respectively (taken from [69]).

The Japanese female facial expression (*JAFFE*) database is a static and grayscale facial expression database which collects six basic facial expressions and the neutral facial expression from 10 female Japanese models [69]. Each subject was requested to pose 3 or 4 times for each of the prototypic facial expressions and once for the neutral

expression. As a result, the *JAFFE* database contains a total of 213 posed facial expression images. The image spatial resolution used for each image is 256×256. In regard to head pose, minor in-plane rotation is often observed while out-of-plane rotation is observed to be minor to none.
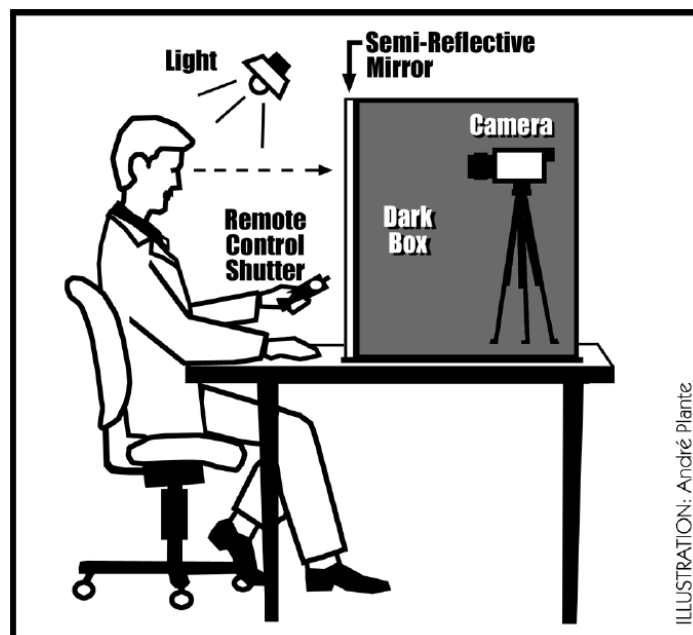
Figure 2.2: An illustration of the environmental setup for data collection for the Japanese female facial expression database (taken from [69]).

The data collection process was carried out in a special conditioned room. The source of illumination was placed above and in front of the subject to ensure the illumination is roughly even across the entire face region. The subject takes a picture of herself while looking at a reflective sheet, which offers the subject the opportunity to judge the quality and the accuracy of the expression executed based on her knowledge of the expression seen in the reflection, to decide which is the best frame that describes the requested expression. It is worth pointing out that a data collection without self-verification is different from one with self-verification because the outlined procedure of the data collection process has enabled the subject to adjust her performance

according to her knowledge of the requested expression. Consequently, the collected images of facial expression in the database will eventually be shifted toward their understanding of the imagery display of the requested facial expression rather than a spontaneous expression they give in day-to-day living. Figure 2.2 briefly illustrates the environmental setup of data collection stage.

Furthermore, the database adopted a semantic rating scheme on the collected samples. A total of 92 Japanese female subjects were requested to inspect and rate samples in the database towards six affective words within a range of 1 to 5, with 1 indicating perceived low similarity and 5 indicating high perceived similarity. Individuals involved in the rating procedure were divided into two groups. One group rated the entire set of database samples and the other group rated all facial expressions in the database except those of the facial expression of *fear*. The reason for this rating setup is that they observed that participants have difficulty in posing *fear*. The final rating score for each image is an average of the rating scores from the two groups, which serves as an indication of the validity and accuracy of the posed expressions. This rating is utilized to label the expression in each image. An image is labelled according to the expression with the highest similarity rating.

The Japanese female facial expression database has been used by many researchers in many publications to evaluate facial expression recognition systems [110], [111], [112], [83], [113].

2.1.2 The extended Cohn-Kanade Database

The extended Cohn-Kanade (*CK+*) database [114] is an extended version of the original Cohn-Kanade database with validated emotional labels and facial action coding for automatic facial expression recognition research. After extension, the *CK+* database records facial actions from a total of 210 adults, who are aged between 18

and 50 years, and with a distribution of 69% female and 31% male. The *CK+* database consists of 81% Euro-American subjects, 13% African-American, and 6% other groups (not specified in [114]). Participants are instructed to pose facial actions (AUs) that are described in facial action coding system (*FACS*) [6], including single facial action and combined facial actions. A detailed list of AUs coded in the *CK+* database is tabulated in Table 2.2. A total of 593 video sequences of posed prototypical facial expressions are collected excluding sequences of facial actions from 123 subjects coded in various numbers of frames (specifically, from 10 to 60 frames). The first and last frames are the onset (neutral) and apex (peak frame) of an expression respectively. 309 in 593 video sequences are validated and labelled as prototypic facial expressions, excluding *contempt* which is not a prototypic facial expressions defined by Ekman [9]. The complete statistics of the collected and labelled expressions are tabulated in Table 2.1. The *CK+* database includes both grayscale and colour facial images of two different resolutions, 640×490 for grayscale images and 640×480 for color images.

Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
No.	45	59	25	69	28	83

Table 2.1: Statistics of labelled prototypic facial expressions contained in the *CK+* facial expression database.

Figure 2.3: Examples of facial expression images contained in *CK+* images (taken from [115]).

AU	Name	N	AU	Name	N	AU	Name	N
1	Inner Brow Raiser	173	13	Cheek Puller	2	25	Lips Part	287
2	Outer Brow Raiser	116	14	Dimpler	29	26	Jaw Drop	48
4	Brow Lowerer	191	15	Lip Corner Depressor	89	27	Mouth Stretch	81
5	Upper Lip Raiser	102	16	Lower Lip Depressor	24	28	Lip Suck	1
6	Cheek Raiser	122	17	Chin Raiser	196	29	Jaw Thrust	1
7	Lid Tightener	119	18	Lip Puckerer	9	31	Jaw Clencher	3
9	Nose Wrinkler	74	20	Lip Stretcher	77	34	Cheek Puff	1
10	Upper Lip Raiser	21	21	Neck Tightener	3	38	Nostril Dilator	29
11	Nasolabial Deepener	33	23	Lip Tightener	59	39	Nostril Compressor	16
12	Lip Corner Puller	111	24	Lip Pressor	57	43	Eyes Closed	9

Table 2.2: The total number of action units (AUs) that is presented on the peak frames in *CK+* database (taken from [114]). *N* represent the number of times the AU is coded.

(a)   (b)

Figure 2.4: Examples of expressive images of the same subject captured at two views, frontal (a) and 30° (b) (taken from [115]).

The data collection of the *CK+* database is set up in an observation room. Two digital cameras are equipped to collect video sequences of facial expressions from a frontal view and at 30° to the subject's right at the same time. Two examples of facial expression images captured at these two camera views are shown in Figure 2.4. The subject is seated while the data collection is carried out. Regarding illumination control,

two different setups are used: either an ambient room light with a high-intensity lamp or two high-intensity lamps with reflective umbrellas. It is worth mentioning that, despite measures being applied to assure a uniform illumination condition, some images in the *CK+* database are overexposed to a minor extent. Regarding head pose variation, in-plane and out-of-plane motion is minor to none.

To validate the emotion labels of the collected facial expression data, interpreted criteria of prototypic facial expressions provided by *FACS* [6] as shown in Table 2.3 are measured on the peak frames of all 593 video sequences. In the end, through a three-step validation process, 309 video sequences have been labelled as representing one of the six prototypic expressions excluding *contempt*.

Emotion	Criteria
Angry	AU23 and AU24 must be present in the AU combination
Disgust	Either AU9 or AU10 must be present
Fear	AU combination of AU1+2+4 must be present, unless AU5 is of intensity E then AU4 can be absent
Happy	AU12 must be present
Sadness	Either AU1+4+15 or 11 must be present. An exception is AU6+15
Surprise	Either AU1+2 or 5 must be present and the intensity of AU5 must not be stronger than B
Contempt	AU14 must be present (either unilateral or bilateral)

Table 2.3: The criteria for interpretation of emotion in terms of facial AUs. The AU identifier number used here is described in Table 2.2.

The extended Cohn-Kanade facial expression database is one of the most popular facial expression databases used in facial expression research. Reporting our system's performance on this database will allow our proposed system easily to be compared to many other state-of-the-art facial expression recognition systems in the field.

2.1.3 Binghamton University 3D facial expression (*BU-3DFE*) Database

The Binghamton University 3D Facial Expression (*BU-3DFE*) database [116] is a 3 dimensional facial expression recognition database, which contains six prototypic facial expressions from 100 subjects who are undergraduate and graduate students at the State University of New York at Binghamton. Each subject was instructed to pose prototypic facial expressions including *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. All six basic facial expressions are recorded in four different intensity levels except neutral, as illustrated in Figure 2.5. As a result, the database consists of 2,500 geometric models and 2500 texture pictures taken from the left and right point of view of the subject. The proportion of male to female subjects in the database is $3:2$. The subjects in the database are drawn from various ethnic groups, including white, black, East-Asian, Middle-east Asian, Hispanic Latino, and others.
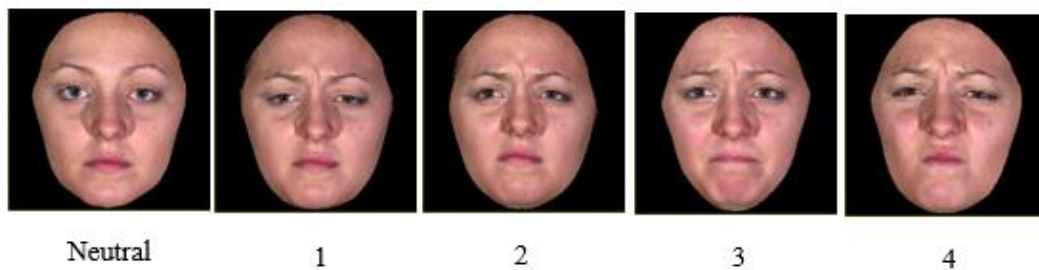
Figure 2.5: Examples of neutral and anger of the same subject in *BU-3DFE* database. Images from left to right are the neutral face, and *anger* in intensity levels of 1 to 4.

To extract the 3D facial expression, the 3dMD digitizer is employed [117]. The detailed setup of the data collection environment is illustrated in Figure 2.6. Six synchronized cameras and two flashing system are deployed to build the system, which are configured as seen in Figure 2.6 [116], [117]. The system generates two texture images in the resolution of 1300×900 pixels and a polygon face model in resolution between 20,000 and 35,000 polygons.
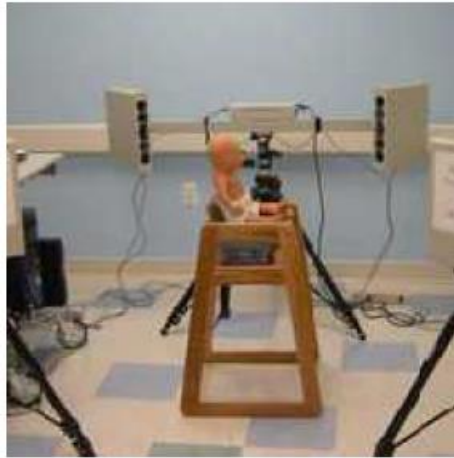
Figure 2.6: Environment setup for the data collection process (taken from [116]).

(a)

(b)

(c)

(d)
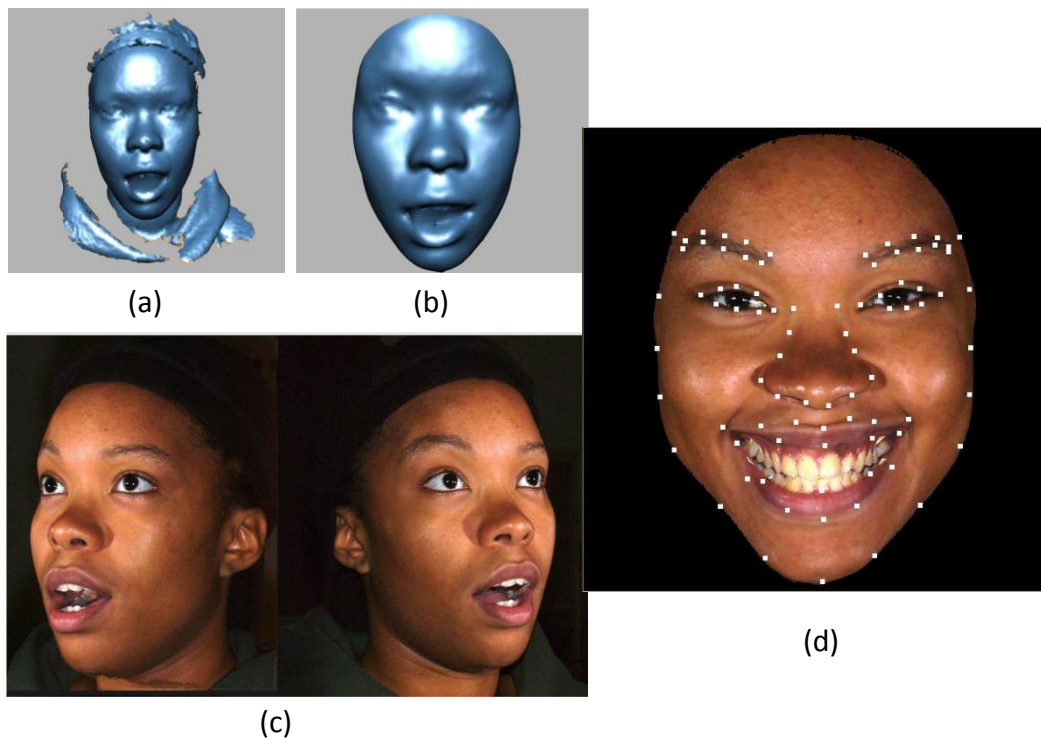
Figure 2.7: Examples of 3D shape models and texture images in the *BU-3DFE* database: (a) raw 3D models; (b) processed 3D models with only face regions; (c) texture images of two views; (d) annotated facial feature points (taken from [116]).

The system automatically processes the captured data and produces a 3D geometric shape of the face along with texture images, which describe the appearance of the surface of the face. After the 3D facial expression data is collected, a processing step is carried out to remove non-face regions contained in the original 3D model and generate a frontal texture image, as demonstrated in Figure 2.7. In addition, a set of marked facial feature points is also provided as illustrated in Figure 2.7 (d).

During a data collection session, the subject, sitting frontal to the 3D face system, is instructed to perform six facial expressions at four different intensity levels based on their own understanding of intensity levels of facial expressions. The intensity levels are described by the instructor as low, middle, high, and highest intensity. Consequently, each subject poses 25 samples of facial expressions, four for each basic facial expressions and 1 for *neutral*.

The validation of the collected data is performed in three steps: first, by the subject him/herself; second, by an observer who is expert in reading facial expression; finally, by a facial expression analysis system. The subject's vote about whether the expression shown is the requested expression is considered as the ground truth of the expression.

2.1.4 Simulation of facial images from *BU-3DFE* database

Since the aim of our research is to devise a 2 dimensional multi-view facial expression recognition system, a large numbers of multi-view facial expression images are required. The *BU-3DFE* database has enabled us to reconstruct the 3D facial expression data from the 3D geometric shape and texture images provided in the database and simulate the required 2D multi-view facial images for our study. To reconstruct the 3D facial expressions, the Simulink 3D Animation Toolbox [118] is utilized throughout the entire simulation process. The simulation process consists of the following steps:

1) First, the 3D face model is loaded and reconstructed by the virtual world viewer. Then a set of 35 viewpoints are created, from which a 2D facial image is captured and saved.

2) After that, the simulated 2D facial image is cropped and resized to 128×96.

3) Finally, the image is converted from RGB to grayscale images, and archived for experiments.
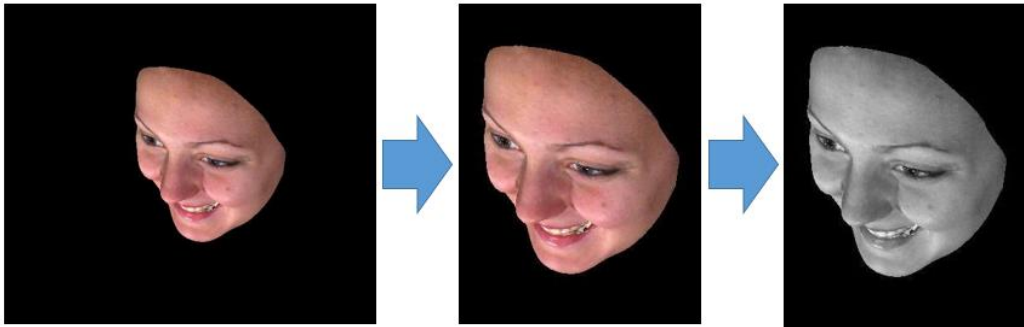
Figure 2.8: An example of the simulation process. From left to right we see the simulated raw image at viewpoint of pan angle of $-30°$ and tilt angle $-30°$, the cropped image, and the archived final image.

2.2 In-house multi-view facial expression database

A suitable facial expression database is an essential tool for researching and developing facial expression recognition systems. Although many facial expression databases has been introduced over the last decades [69], [114], [116], [119], [120], [121], none of them can allow researchers to thoroughly investigate the influence of different views on the recognition of all prototypic facial expressions. Therefore, it

was considered necessary to create a real multi-view database for more detailed exploration of facial expression recognition from various poses.

2.2.1 General information

Expressions	Six prototypic facial expressions
Pan views	−90°, −60°, −30°, 0°, 30°, 60°, 90°
Illumination	One ambient room light
Stimuli	Oral instruction
Videos	252 video sequences
Subjects	6
Age	25-35
Ethnics	East Asian, European
Accessory	Glasses and scarf
Resolution	1280 × 720
Fame/second	29 fps

Table 2.4: General information and statistics of the in-house database

Our in-house multi-view facial expression database was designed for exploring, devising and evaluating the preliminary design of our proposed facial expression recognition system. This database contains six prototypic facial expressions from 6 participants who are postgraduate students at the University of Kent. For each subject, six basic facial expressions, including *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*, are collected at 7 different pan views. Consequently, a total of 252 raw video sequences of facial expressions are collected from all subjects in two sessions with an interval of 1 week between sessions. Each sequence starts from a neutral face to apex of the expression, and then back to neutral face again. The ethnic groups in the database are East-Asian and European, and participants are aged between 25 and 35.

Table 2.4 summarizes the general information of the in-house multi-view facial expression database.

2.2.2 Environmental setup

The data collection took place in a small laboratory room equipped with a Nikon D7000 camera. The digital camera is set up in front of the subject at a fixed distance of 1.2 meters. An office chair is placed in the middle of the scene for the subject. At the beginning of each data collection session, the height of the camera is adjusted to keep the tilt angles to the minimal degree. To acquire multi-view facial expressions, a single camera sequential acquisition scheme is utilized to acquire facial expression from each of the pan views respectively. Six points are marked down on the wall for every pan view except the frontal view, for the subject to look at while posing their facial expressions. Regarding illumination, one ambient room light is used, which is mounted in the ceiling above the camera and in front of the subject. A demonstration of the environmental setup is shown in Figure 2.9.
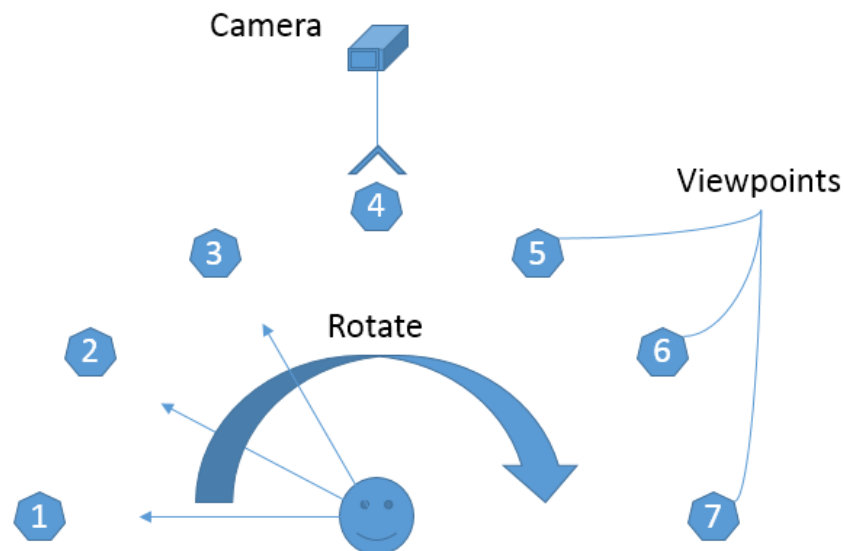
Figure 2.9: Environmental setup for the data collection process.

2.2.3 Data collection

During the data collection process, the subject is seated while performing the required expressions. As demonstrated in Figure 2.9, facial expressions from 7 different viewpoints are acquired successively from left to right. At each view, each subject is requested to pose the expression based on their own understanding of the given affective adjectives comprising *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. These expressions are posed one after another and one expression at a time at each viewpoints. The data acquisition started from the left profile view, which is labelled as 1 as shown in Figure 2.9, and then facial expressions from other viewpoints are collected in clockwise order. The data acquisition takes place in two sessions with an interval of 1 week between sessions to introduce intra-class variation and subject variation. The same data collection protocol is followed in all data collection sessions.

2.2.4 Validation of acquired facial expressions

The validation of facial expression data is completed in one step. Each subject reviews his/her facial expressions after capture. If the subject confirms the expression is not the requested facial expression, the acquired facial expression is removed, otherwise it is archived.

3.2.5 Quality control of the data

The subjects are allowed to wear accessories throughout the data collection session. One third of subjects wore glasses, and one sixth of subjects wore a headscarf. As a result, half of the subjects in the database wore accessories, which would bring substantial difficulties in recognition of facial expressions on the database. It should be also noted that some acquired facial expression data are slightly blurred due to exaggerated movement of the head during the capture. To sum up, facial expressions possessing the following properties are removed: first, off-focus; second, video

sequences that do not demonstrate the requested facial expression as confirmed by the subject. As a result, the final archived database contains 196 video sequences.

2.2.5 Facial expression data

As validation of facial expressions and data quality control, 196 video sequences are archived in the database. In addition, the peak frame of facial expression in each video sequence is captured for the purpose of devising a static facial expression recognition system. Some examples of captured peak frames of facial expressions are provided in Figure 2.10.

Figure 2.10: Some examples of captured facial expressions in our in-house database: (a) *anger*; (b) *happiness*; (c) *sadness*; (d) *neutral*.

2.2.6 Advantage and disadvantage of the database

Our in-house database has several pros and cons. First, the size of the database is small. It is difficult to devise a reliable system from this amount of data. The second point is

that a large proportion of the participants wear accessories which make it difficult for devising facial expression recognition solutions without accessories but good for testing a system's robustness. Third, there are shadows on the performer's face due to insufficient illumination compensation. These properties introduce considerable uncertainty into the evaluation of the performance of the facial expression recognition system built based on this database, they could also be an advantage because issues including illumination and accessory inclusion is common in practical application scenarios, which make this database feasible for evaluating the robustness and practical performance of a facial expression recognition system. Most importantly, however, is that - while acknowledging the limitations - this database allows us to examine, at least in a preliminary way, important image types which could not otherwise be investigated in the reported study.

2.3 Face detection

Figure 2.11: An example of a detected face region using the Viola and Jones's face detector on an image from the *CK+* database [114].

As described previously in Chapter 1, face detection is an essential part of a facial expression recognition system as it locates the initial region of interest in the source image. In this study, the well-established Viola and Jones's face detection algorithm [33] is employed because of its effectiveness and its efficiency in detecting faces in real time. A toolbox implementing the Viola and Jones's face detector by Kroon is utilized in this study [122]. Figure 2.11 demonstrates a face region as detected by this face detector.

2.4 Classification using support vector machines

In this study, the multi-class support vector machine (*SVM*) classifier is deployed to classify facial expressions [123]. The support vector machine classifier is an advanced and well-established classifier, which has been deployed to solve a large variety of classification problems, with particularly promising results having been reported in many related facial expression recognition research studies [93], [124], [67], [54], [125]. Therefore, the support vector machine classifier is utilized throughout this research. The multi-class *SVM* classification is achieved by adopting the one-vs-all classification strategy suggested by Crammer and Singer [126], which turns a multi-class classification problem into a constrained optimization problem and allows the problem to be solved more efficiently. A toolbox which implements and optimizes the original support vector machine for large scale applications is utilized [127].

For all the results reported in this thesis, the classification accuracy that is reported is an average of the classification accuracy of ten repetitions of stratified 10-fold cross validations. Kohavi [128] studied the methodology of cross-validation and suggested to use stratified 10-fold cross validation for model selection because it gives a better estimation of the classification accuracy. A large partition that is more than 10 folds gives rise to bias while a small partition increases variance. Therefore, the recommended cross-validation strategy is deployed in this study for reporting the classification accuracy.

2.5 Conclusion

In this chapter, the primary databases that are utilized in devising and testing our multi-view facial expression recognition systems throughout this study are described in detail. In addition, we have also presented a novel facial expression database, which is designed and compiled in our laboratory specifically for multi-view facial expression research. Finally, general techniques and toolboxes that have been adopted for simulation of facial expression images from a 3D model, for face region detection, and for the classification of facial expressions have been explained.

In the next chapter, a general review of state-of-the-art texture-based multi-view facial expression recognition systems is elaborated, and then our novel facial expression recognition approach using local ternary pattern and multi-scale local ternary pattern operators are presented, for which outstanding performance has been achieved.

Chapter 3

Local ternary pattern based universal facial expression recognition

In this chapter, a review of some state-of-the-art multi-view facial expression recognition systems and analysis of the advantages and disadvantages of these systems is at first presented. Then, an investigation of possible different configurations of local ternary pattern operator in the application of universal multi-view facial expression recognition is elaborated. In addition, the original local ternary pattern operator is extended to operate at multiple scales for which, when employed for constructing multi-view facial expression recognition system, a state-of-the-art performance is achieved.

In section 3.1, the motivation and general background of this research is described. In section 3.2, related research publications and state-of-the-art multi-view facial expression recognition systems are presented and analysed. Section 3.3 describes the structure of the proposed novel multi-view facial expression recognition system. Section 3.4 describes the texture feature representation method employed in this study. In section 3.5, the block based local feature extraction technique is described. Section 3.6 explains the feature selection algorithm adopted in this research. In section 3.7, the detail experimental set up and analysis of results is presented. Finally, in section 3.8, the contribution of this chapter's content is summarized.

3.1 Introduction

Over the past few decades, many research studies have been developed which address the issue of automating the facial expression recognition process. More recently, automatic facial expression recognition has been increasingly researched due to its potential applications in the area of human machine interaction and so on. However, the problem relating to automatic recognition of facial expression in uncontrolled environments or with various head pose is still not generally well understood and has not been fully addressed by the research community. Although some multi-view facial expression recognition systems have been proposed by the pioneers in the field, the performance of these systems have not yet fully resolved this research problem in order to facilitate the use of facial expression recognition system in practical applications. In this study, a novel universal multi-view facial expression recognition system is introduced using local ternary patterns and multi-scale local ternary patterns which is extended using a similar approach as the multi-scale local binary pattern approach [129]. This proposed novel universal facial expression recognition system has achieved state-of-the-art performance in classification accuracy even when used with a great range of head poses, including 7 pan angles and 5 tilt angles during image capture. In addition, a series of experiments are also carried out to investigate the influence of various configurations of the local ternary pattern operator in constructing a robust and efficient feature representation for universal multi-view facial expression recognition. The detailed experimental setup and results are also reported in this chapter.

In recent years, in order to obtain an efficient and robust solution for multi-view facial expression recognition, various texture descriptors have been developed, introduced, and employed to generate a texture feature representation for facial expression images encompassing different viewing angles, including local binary patterns [24], Gabor features [130], and so on. Different from geometric based facial expression recognition algorithms or a combined geometric and appearance based approaches, the texture based facial expression recognition algorithm adopts a holistic or local texture feature

extraction process that makes texture based facial expression recognition system more straightforward to implement. The state-of-the-art texture descriptors include scale invariant feature transform ($SIFT$) [131], local binary pattern (LBP) [132], Gabor features [5], speed up robust feature ($SURF$) [37], gradient location and orientation histogram ($GLOH$) [74], histogram of oriented gradient (HOG) [75], discrete cosine transform (DCT) [76], and so on. Among them, the scale invariant feature transform and local binary pattern feature have been widely adopted for facial expression recognition purposes, and promising system performance has been reported for frontal, near frontal and view- dependent facial expression recognition [133], [134], [135],[136].

Although intensive research has been carried out to investigate the multi-view facial expression recognition task, there are still practical problems in applications of these laboratory-proven solutions because real recognition scenarios tend to diverge from ideal experimental conditions. The problematic issues generally fall into two categories: internal issues and external problems. The internal issues of a system derive from the fact that each individual component of a facial expression recognition system is a potential error source, and can contribute to the overall error rate of the whole system. A typical example is that the raw data source exerts a significant influence on both the performance and usability of facial expression recognition systems. The recognition model and mechanism is first built based on an exploration of the raw data source, which requires that the original data must be either designed and populated as close to the real application scenario as possible, or must be compensated so that it corresponds seamlessly to the real day-to-day data. Otherwise, the derived solution will either be biased or will shift far away from the accuracy obtained in the ideal laboratory setup. External environmental issues, including illumination, head pose, and distance away from camera, also affect the recognition accuracy and usability of a facial expression recognition system.

Among the issues mentioned above, out-of-plane angle (i.e. head pose presented in

the facial images) is a crucial factor, not only because it is inevitable in some application scenarios but also because it is an important element to establish and improve system usability. It is acknowledged that head pose can significantly affect the performance of the face recognition task, and this is still a challenging research issue even though enormous efforts have been made to resolve the problem [91]. For applications in constrained environment, certain restrictions can be applied (or user instructions given) to control the head pose for some applications, but elimination of head pose in practice is infeasible, and consequently can affect the performance and general usability of a system. For applications in unconstrained environments, out of plane angles carry a more significant influence over the performance of a facial expression recognition system. For example, a large variation of head pose can exclude a large part of the face from the captured facial image and hence cause a geometric model based feature extractor to fail to generate a feature or extract an invalid feature. Therefore, it is necessary for a facial expression recognition system to accommodate various head poses to achieve better usability and wider application with higher robustness, especially in an unconstrained application scenario. In order to resolve the issues arising from various pan and tilt angles in a constrained or unconstrained environment, researchers have developed several recognition mechanisms, which can be grouped into three categories:

- 2D view dependent facial expression recognition systems: these create a parallel recognition unit for each defined view and, depending on the result of the head pose estimation a specific classification unit is activated to extract the feature representation and classify the presented expression.

- 2D universal facial expression recognition systems: this type of system classifies facial expressions collected from various views in the same way. Feature extraction and classification are completed universally, and therefore the usability of such systems in terms of pose handling is greatly improved.

- 3D facial expression recognition systems: these have a high demand in hardware and software, such as computation power, system memory, and optical components, which therefore makes the system expensive to deploy. To reduce the computational complexity, usually a reduced 3D statistical model is used to approximate the geometric shape of the actual facial surface, and obtained parameters of the 3D shape model and textures are employed as feature representations for 3D facial expression recognition.

Among these three categories of recognition mechanisms, 2D universal facial expression recognition systems thoroughly investigate the characteristic of facial expression data from all defined views, and deliver a universal solution for the multi-view facial expression recognition task. Although the recognition is generally reported to be poor, the completeness and usability of the solution can offset its disadvantages. The view dependent systems tend to deliver a solution slightly better than 2D universal facial expression recognition in terms of classification accuracy, but greatly rely on the measure from a pose estimator, which restricts its overall performance and capability to handle facial expression recognition from arbitrary views. 3D facial model systems generally give the highest system performance, but they have a higher demand on the system hardware, such as computational power, system memory, and optical components, and therefore make this kind of system more expensive to deploy in practice. Compared with the other mechanisms, universal 2D facial expression recognition systems are inexpensive and straightforward in system design and setup.

In this chapter, a novel universal multi-view facial expression recognition system based on local ternary pattern and the extended multi-scale local ternary patterns will be presented along with a comprehensive analysis of experimental results.

3.2 Related work

Texture analysis has found application in many fields ranging across object identification, image region classification, image segmentation, content-based image, and information retrieval, medical image analysis, and so on. With many years of research and development, increasingly more texture operators have been deployed to analyse multi-view facial expressions, and some of these have achieved state-of-the-art performance, including the local binary pattern operator and other variants derived from it [130], [137], [73], [138], [125], histogram of oriented gradient [139], discrete cosine transform [76], Gabor filter [140], and scale invariant feature transform [76], [133], [135], [138].

3.2.1 2D view dependent and 3D facial expression recognition systems

Many researchers have turned to local descriptors to find solutions for the issues raised by head pose variance. One of the most influential methods is based on scale invariant feature transform descriptor, which gives promising results in handling head pose variance. Berretti et al. [141] introduced a solution based on *SIFT* features extracted from around 112 facial landmarks on the 3D depth image. Rudovic [142] introduced multiple regression models to classify the facial expression image at various head poses. Eleftheriadis [88] derived an approach based on a shared Gaussian process latent variable model (*SGPLVM*). Hesse et al. [76] proposed a system using a combination of facial landmark coordinates and discrete cosine transform (*DCT*) feature extracted around facial landmarks.

Soyel and Demirel [138] introduced an affine transform-based pose invariant *SIFT* (*Π-SIFT*), which consists of five steps as illustrated in Figure 3.1:

1) First of all, dense *SIFT* features are extracted from image pairs;
2) Secondly, the best matched points with the extracted *SIFT* features in two

images are obtained using Kullback Leibler Divergence (*KLD*) based *SIFT* matching algorithm [143];

3) Then the affine transformation is estimated based on best matched points found in *KLD* based *SIFT* matching algorithm using singular value decomposition;

4) After that, *SIFT* points are regrouped into regions in preparation for grid based *SIFT* matching;

5) Finally, the accumulated regional measure of similarity is obtained from the WMV classifier [75], and then weighted by the grid based confidence coefficients to form a final score for all expressions.

(a)   (b)
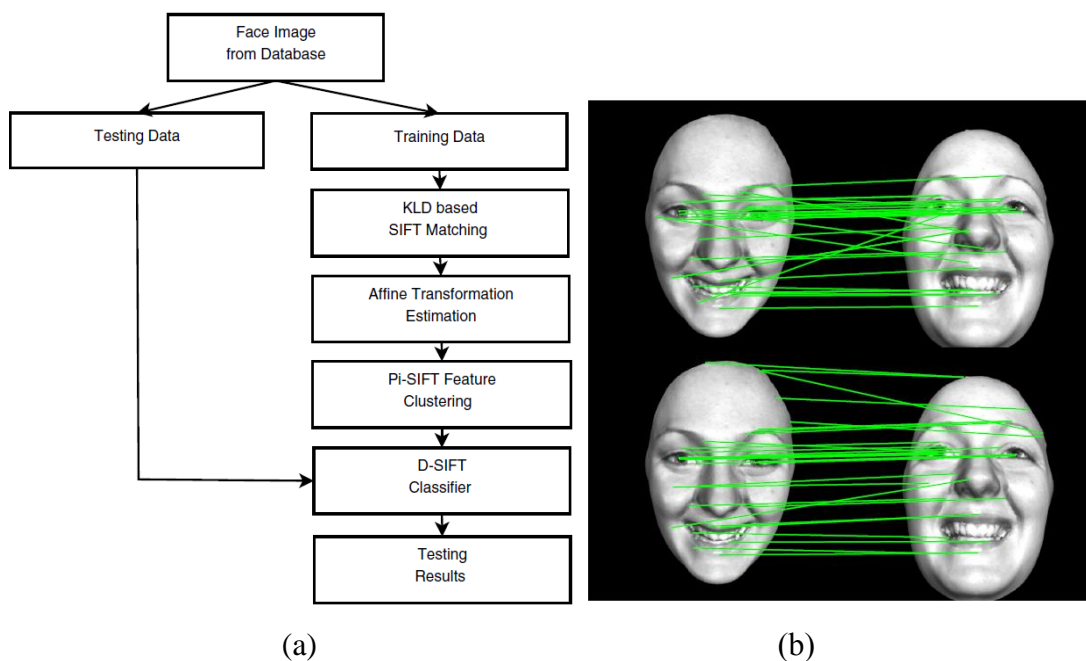
Figure 3.1: (a) The proposed multi-view facial expression recognition system by Soyel and Demirel's is illustrated; (b) an image illustrating the matching process of two correspondent $\prod$-*SIFT* features (taken from [138]).

Their method approached the multiple pose issue by deploying the affine transformation estimated based on the scale invariant transform features, and they claim that the performance of the $\prod$-*SIFT* based multi-view facial expression

recognition system outperformed the regular *SIFT* and *D-SIFT* feature representations. However, this system is computationally expensive to implement, and relies on affine transformation estimation for an accurate head pose estimation. Besides, they did not present a one versus all classification accuracy, which is the de facto measurement for a facial expression recognition system's performance.

3.2.2 Universal multi-view facial expression recognition systems

Other researchers have attempted to resolve the issue by deploying dense feature extraction in the design of facial expression recognition systems, such as dense scale invariant feature transform and so on, which has delivered an encouraging performance in dealing with pose variation in the input facial images.

Zheng et al. proposed another "universal" approach for multi-view facial expression using the regional covariance matrix (*RCM*) and a Bayes discriminant analysis via Gaussian mixture model (*BDA/GMM*) to learn the best representation for facial expression images from arbitrary views [144], [145]. First, dense *SIFT* features are extracted from the identified facial region. As the densely extract *SIFT* feature is redundant, they adopted the proposed *BDA/GMM* algorithm (which is explained thoroughly in [145]) to reduce the dimensionality of extracted *SIFT* features to search for the most discriminant features. Based on the reduced *SIFT* features, the regional covariance matrix is calculated. The overall performance of their proposed system has been examined on the *BU-3DFE* database with the rendered facial expression images of 35 various views of a combination of pan and tilt angles. The overall performance of their system is an accuracy of 68.28%.

Usman et al. [146] introduced a multi-view facial expression system in a "universal" approach which takes generic sparse coding features and recognizes the expression

based on a "universal trained" classifier (i.e. a classifier that is trained with feature representations generated from images of all views). In addition, they introduced a mid-level feature representation (a so-called generic coding feature), which was generated in three phrases (see Figure 3.2):

1) Firstly, a dense feature extraction is carried out to extract *SIFT* features.

2) Secondly, these generated features are encoded using sparse coding [147] to keep the distinctiveness of previous coding.

3) Finally, a spatial feature pooling algorithm is applied to construct a complete representation for the image.
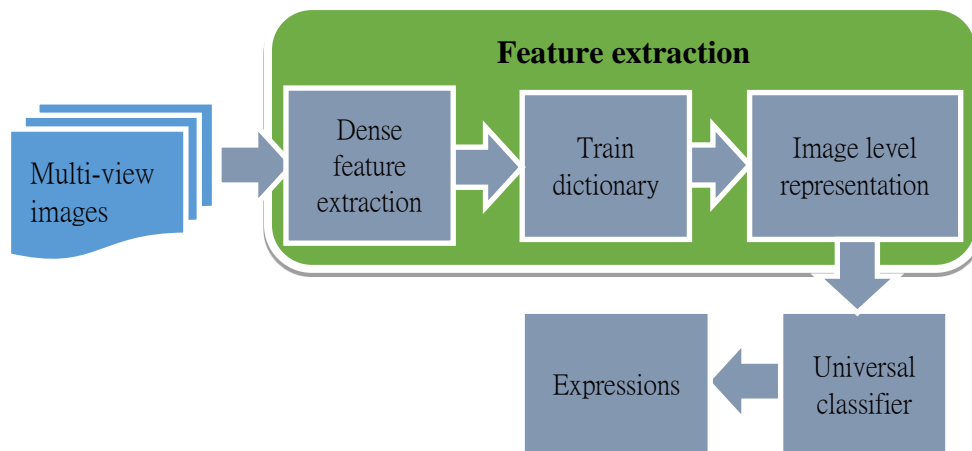
Figure 3.2: System structure using a generic sparse coding feature.

Their system has obtained results on the *BU-3DFE* database with an overall accuracy of 69.1% on facial images of all intensity levels, which is rendered with 3D models and texture images, and 76.1% on the highest intensity facial images only.

Later, in another study, Usman et al. proposed maximum margin Gaussian mixture model (*GMM*) learning for a supervised soft vector quantization (*SSVQ*) for multi-view facial expression recognition. Firstly, they extract the *SIFT* feature vectors using a sliding window of 16×16 pixels with 3 pixels shift across the entire facial regions, and to reduce the dimensionality of the resulting feature vector, the extracted *SIFT* feature vector is reduced from 128 to 70 in length using principal component analysis (*PCA*) [70]. Then, they utilized the expectation maximization algorithm to estimate the initial parameters for the *GMM* model, and restrict the components to 1024 for each model so as to reduce the computational complexity. Finally, supervised soft vector quantization is applied.

Based on 21,000 facial expression images created from data contained in the *BU-3DFE* database, they report an average accuracy of 76.34% using *SSVQ+SPM* features over 35 views including 7 pan angles and 5 tilt angles of the highest intensity level.

These two systems proposed by Usman have achieve an accuracy of 76% and accommodate 35 different views, but the overall performance of their proposed system is limited, and classification accuracy of expressions at low intensity was not reported, which is a crucial evaluation for practical application as spontaneous facial expression is less pronounced than posed expressions.

In our study, the objective of this research is to find a universal multi-view facial expression recognition solution by exploring the use of the local ternary patterns operator, level of difference pattern operator, and other state-of-the-art texture descriptors.

3.3 The proposed universal multi-view facial expression recognition system

In this section, the general structure of our proposed multi-view facial expression recognition system and its key components are elaborated.

The general framework for our proposed system can be decomposed into three distinct and separate stages, comprising the data acquisition stage, the feature extraction stage, and the classification stage, as illustrated in Figure 3.3.
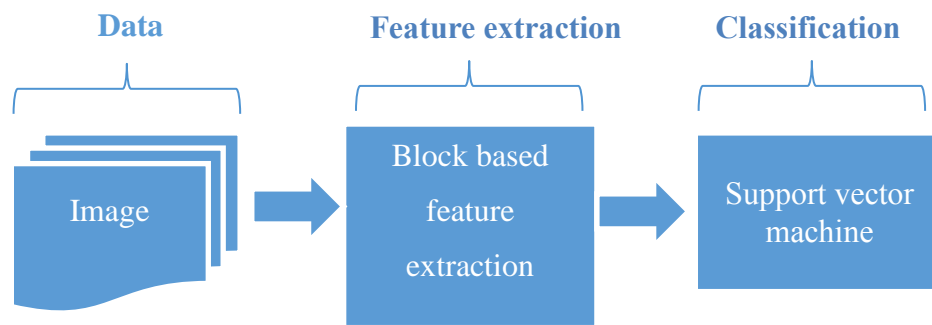
Figure 3.3: The sequence of three stages in the proposed system

1) As shown in Figure 3.3, the data acquisition stage is the first stage of our proposed automatic facial expression recognition system, which acquires a single facial expression image for processing. The detailed description of the data and pre-processing is presented in Section 3.7.1 and 3.7.2.

2) At the feature extraction stage, various local ternary pattern operators are applied on the image to extract local ternary pattern (*LTP*) features using a block based feature extraction method, which is explained in detail in Section 3.4. To reduce the dimensionality of the final representation, the F-score feature selection algorithm [148] is adopted to select the most distinctive features from the

available local ternary patterns features. That is described in Section 3.5.

3) At the classification stage, the pre-trained universal support vector machine classifier is employed to classify the input expressive facial image into one of the prototypic expression categories (Section 2.4 presents the detailed information about the support vector machine classifier).

3.4 Feature extraction and selection

3.4.1 Block based feature extraction

This section will describe in detail the procedures and techniques adopted to extract a block based feature representation for the proposed multi-view facial expression recognition system.

To obtain a more reliable, stable, and robust feature representation, block based feature extraction has been adopted by many precursors for both face and facial expression recognition. Shan et al utilize a block based local binary pattern extractor for facial expression recognition in their facial expression research [130], [137], [73]. In their experiments, they adopt a block size of 18 by 21 pixels which divides an image into $6 \times 7 = 42$ blocks in successive studies of facial expression recognition with the Cohn-Kanade Facial Expression Database [115]. Manikantan proposed a block based discrete cosine transform feature extractor for face recognition [149]. They adopt a block partition of 8×8=64, and achieve a promising recognition rate on four different face databases including ORL, UMIT, Yale B and FERET. Yu et al. adopted a grid-based Weber Local descriptor to learn facial expression from Web images [83] utilizing a 5 by 5 partition, and verified their algorithm across a variety of databases, including images from the Web, *BU-3DFE* database, *JAFFE* database, and Cohn-Kanade database. Tong [150] tested a set of block partition options for local gradient coding based feature extractor for facial expression recognition, and their exhaustive experiments have led them to choose an 8 by 8 image partition.

A questioned image is firstly partitioned into small image blocks, which could be of various sizes, and then a local feature descriptor is applied on each image block. The raw features derived from the blocks are fused into a single representation according to a pre-defined fusion scheme. After fusion, a complete feature representation is finalized. A block-based texture feature extraction would provide a finer feature representation since it extracts more feature details from an image, but at the same time increases the complexity of the generated representation in terms of dimensionality.
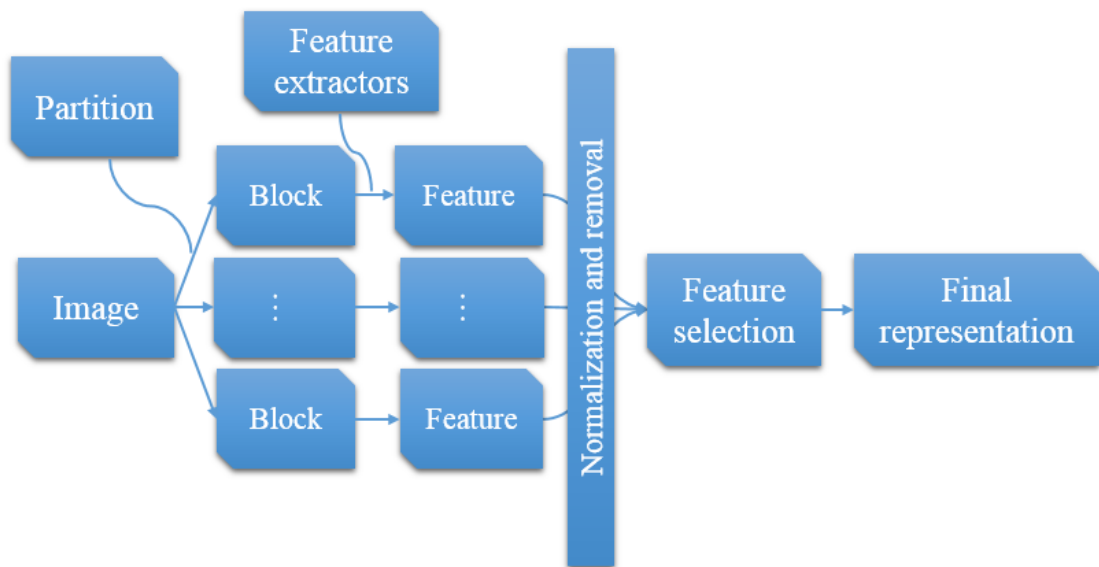
Figure 3.4: An illustration of the detailed feature extraction procedures.

Because of its success and efficiency in improving a texture based feature descriptor, the block-based feature extraction is employed in our proposed system. As illustrated in Figure 3.4, first, a facial image is partitioned into $6 \times 4 = 24$ blocks and each block is of 21 by 24 pixels in size. Secondly, the local ternary pattern operators, LTP or

LTP^{ms} (multi-scale local ternary pattern) operators, are applied respectively on each block. After that, a normalization is performed using Equation 3.1.

$$F = \frac{X - u}{\sigma} \qquad 3.1$$

X is the complete feature; u and σ are the mean and standard deviation of X respectively.
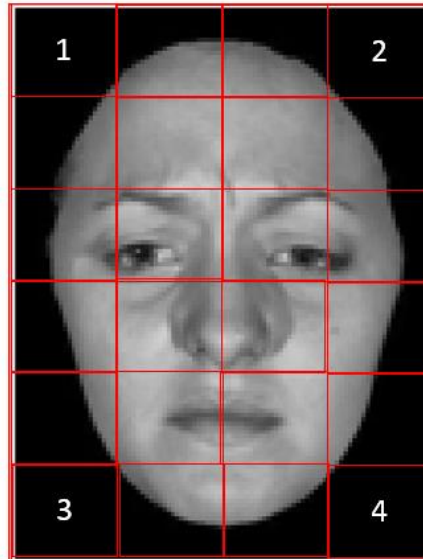
Figure 3.5: An illustration of the blocks marked as 1, 2, 3, and 4.

Third, before the feature selection takes place, a feature removal is performed to remove the features extracted from blocks numbered 1, 2, 3, and 4 (as shown in Figure 3.5), because by inspection of the facial image data, these blocks are found to be dominated by non-facial regions. Finally, the feature selection algorithm, which is described in Section 3.5, is engaged to select the most discriminative feature from the entire feature set to form the final representation. There are two main criteria for selection of this block size. The first is that a block size that is fine and generates a

manageable feature dimensionality will be selected considering the hardware requirement and computation cost for the scale of experiment we will be running. A block size, which is too small, will result in excessively large numbers of features, which would exceed our hardware limit. The other is that a block size that delivers optimal performance in terms of classification accuracy will be selected.

3.4.2 Feature selection

Feature selection is an important stage in the context of machine learning because it reduces the complexity of the classifier and speeds up the operational time of the entire system. In our case, the goal of feature selection is to select the most relevant features for the universal multi-view facial expression recognition task and avoid redundant features that do not possess descriptive information for prototypical facial expressions. The impact of the feature selection becomes more significant when the available feature set is excessively large and sparse. Various feature selection techniques have been adopted to reduce the enormous dimensionality of the feature representation introduced by either a densely extracted texture descriptor or variants of one texture descriptor tuned with different parameters [134], [83], [151], [152], [153]. Taking into account both the influence of dimensionality and the availability of experimental hardware, in this study, a feature selection technique is applied as described below to select the most significant and informative features for representation for a universal multi-view facial expression recognition system.

Specifically, the F-score feature selection technique for feature filtering is adopted in this study. F-score [148] was introduced to measure the difference of two classes. A feature representation is denoted as $X_i, where\ i = 1, \cdots, m$ represents the number of observations. If the two classes are denoted as a and b, then the F-score for the jth feature can be mathematically described as in Equation 3.2:

$$F(j) \equiv \frac{\left(\bar{x}_a - \bar{x}_j\right)^2 + \left(\bar{x}_b - \bar{x}_j\right)^2}{\frac{1}{n_a}\sum_{k=1}^{a}(x_{k,j}-\bar{x}_a)^2 + \frac{1}{n_b}\sum_{k=1}^{b}(x_{k,j}-\bar{x}_b)^2} \qquad \textbf{3.2}$$

In the above expression, $\bar{x}_j$ denotes the average of jth features of X, X_j; $\bar{x}_a$ and $\bar{x}_b$ represent for the average of X_j belongs to class a and class b respectively. $x_{k,j}^a$ and $x_{k,j}^b$ respectively denotes j^{th} feature of the kth observation of class a and b. The numerator describes the distinctiveness of jth feature between two classes, while the denominator describes the variance of the two classes when it is described by the jth feature.

In this study, the F-score criterion is adopted as a filter. For each feature, the F-score is calculated using Equation 3.2, and the features are sorted into descending order. By defining t, the total numbers of feature to select, features with the highest F-score are chosen.

3.5 Local ternary pattern and local binary patterns

The local ternary pattern (LTP) operator is an improved version of the local binary pattern operator which was originally introduced to reduce errors caused by the near-uniform noise in the image [154]. In order to explain the local ternary pattern, in this section, a detailed description of original local binary pattern is given, and then the local ternary patterns employed in this study are presented. In addition, a novel extended version of local ternary pattern operators is also introduced, which enables the original local ternary patterns operators to extract LTP features from multiple scales. The complete explanation of the multi-scale local ternary pattern operator is also included in this section.

3.5.1 Local binary pattern and its variants

3.5.1.1 Basic local binary pattern operator

Ojala, el al [155] established the basic concept of the local binary pattern operator for texture analysis. The original concept of the local binary pattern is to binarize neighbouring pixels of a 3×3 block with intensity value of its central pixels, and concatenates the resulting binary values into an 8-bit binary number as a representation of the micro texture pattern observed in this block. If the neighbouring pixel's value is greater or equal to value of the centre pixel, the pixel is coded as 1, otherwise as 0. The resulting binary code chain is concatenated circularly to form an 8-bit binary code. Each local binary pattern represents a local primitive, which is also called a micro texton. Examples include, spot, flat region, line end, edges, and so on. Figure 3.6 provides some examples of the typical local primitives. The total number of coded variants of micro textons detected by the local binary pattern operator is $2^8 = 256$.
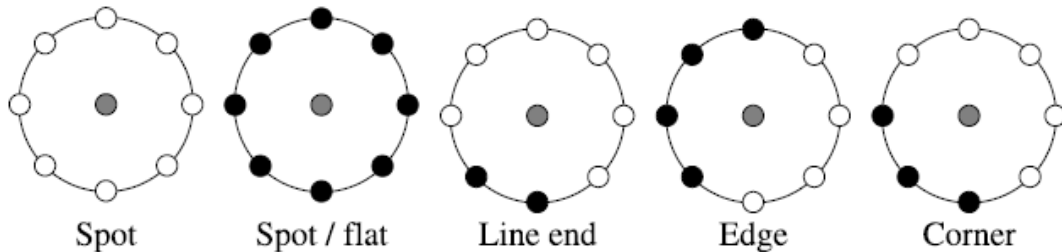
Figure 3.6: Some examples of typical primitives extracted by the local binary pattern operator. (Taken from [129])

Through thresholding and considering only the sign of difference of central pixels with its neighbours, the local binary pattern operator achieves scaling invariance of grey scale, and the local binary patterns are also theoretically robust to monotonic transformation of grey scale.

Mathematically, the basic local binary pattern operator can be summarized in Equations 3.3 and 3.4:

$$G(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad\qquad 3.3$$

In the above expression, $G(x)$ is a thresholding function.

$$LBP = \sum_{i=1}^{8} G(p_i - c)2^i \qquad\qquad 3.4$$

In the above expression p_i is the value of a neighbouring pixel; i is the index of the neighbouring pixels; c is value of the central pixel, and LBP is a decimal representation of the local binary pattern with a range of 0 to 256.
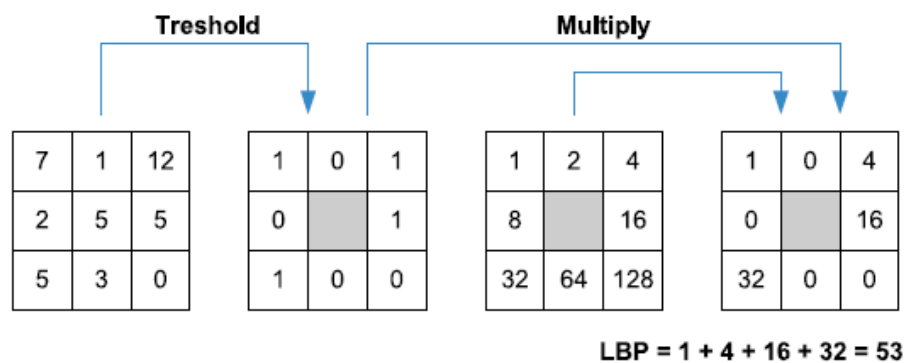
Figure 3.7: Thresholding and coding of local binary patterns (Taken from [156].

Figure 3.7 illustrates a complete calculation process of the basic local binary pattern. A basic local binary pattern feature representation is an occurrence histogram of local

85

binary patterns in regions of interest.

In this study, the basic local binary pattern operator is not used due to the large dimensionality and redundancy introduced by the operator and the limitation of our computer hardware. The detail of the basic local binary pattern is included here as fundamental background for understanding its variants and the local ternary pattern and its variants.

3.5.1.2 Uniform local binary pattern

Due to the simplicity and efficiency of the local binary pattern operator, the methodology of local binary patterns has been intensively researched, expanded, and improved to resolve other complicated computer vision and pattern recognition tasks.
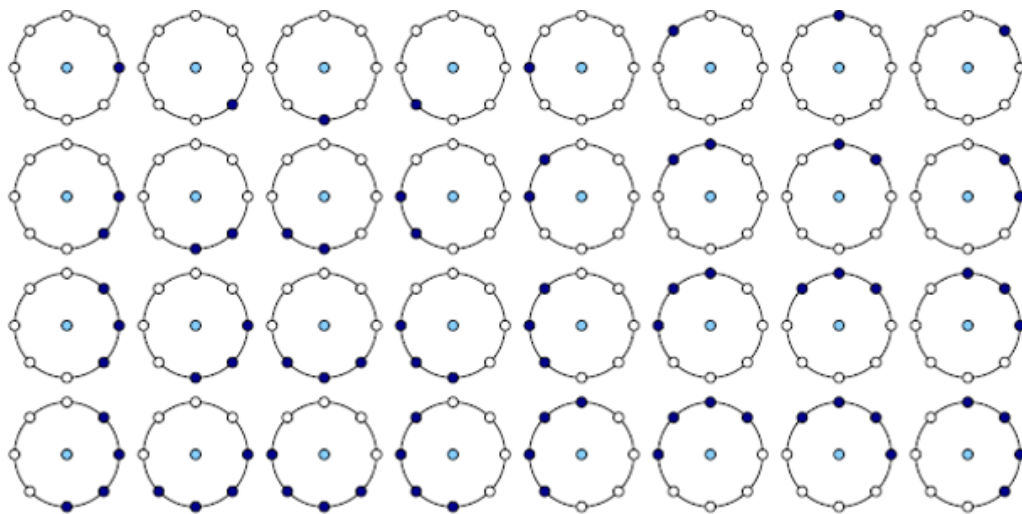
Figure 3.8: An illustration of 36 uniform local binary patterns [129], [140].

In 2002, Ojala, et al. [132] introduced the uniform local binary pattern. The uniform local binary pattern is a subset of local binary patterns that will not be affected by the in- plane rotation of the image patch. In other words, they are circularly symmetrical in appearance and at the same time comprise the majority characteristic of a complete

set of the basic local binary patterns. A local binary pattern is considered 'uniform' if the transition from 0 to 1 and 1 to 0 in its binary code is less than two. For example, 01000000 and 00111000 are uniform local binary pattern with 2 transitions, while 01011101 is not uniform with more than 2 transitions in its binary code. Figure 3.8 displays 32 examples of the uniform local binary patterns operator with sampling radius of 1 and 8 sampling points. Compared with the basic local binary pattern, the uniform local binary pattern reduces the dimensionality of feature representation from the 256 to 59.

When compared with the basic local binary operator, the uniform local binary pattern operator detects a reduced set of variants of micro-textons which capture 90.6% of all primitives detected by a basic local binary operator at radius of 1 with sampling rate of 8, and 85.2% at radius of 2 in an examination of local binary pattern in facial image [157]. It was also noted that uniform local binary patterns tend to be more reliable and robust against noise and statistically yield better performance in many applications [129].

The uniform local binary pattern operator works in the following steps:

1. Basic local binary patterns are extracted using Equation 3.3 and 3.4.

2. Uniform pattern mapping is applied, which assigns each uniform pattern with new labels and non-uniform local binary patterns with the same label, which sums up all other miscellaneous patterns.

3. The resulting pattern is an occurrence histogram with 58 uniform pattern bins and 1 non-uniform pattern bin, and altogether they form a uniform local binary pattern feature representation with 59 dimensions.

3.5.1.3 Multi-scale local binary pattern

In order to address the uneven sampling rate and extend the robustness of local binary

patterns, the original local binary pattern operator is extended to extract local primitives from sampling circles of any radius with evenly spaced sampling points on the diameter, which is designated as multi-scale local binary pattern [132], [129]. By adopting multi-scale local binary pattern operators, each pixel in the region of interest contributes to various extraction processes of multi-scale local binary operators. The extended local binary operator is summarized in Equations 3.5 and 3.6:

Given that (x_c, y_c) is the coordinate of the centre of the sampling circle, and the coordinates of the even spaced points on the sampling circle of radius r are defined as x_i and y_i, then x_i and y_i can be calculated via the following equations,

$$x_i = x_c + r * \cos\left(\frac{2\pi i}{p}\right) \qquad\qquad 3.5$$

$$y_i = y_c - r * \sin\left(\frac{2\pi i}{p}\right) \qquad\qquad 3.6$$

With these equations, the basic local binary pattern operator can be extended to extract local binary patterns at r radius with p sampling points. Figure 3.9 shows five local binary operators at different scale.
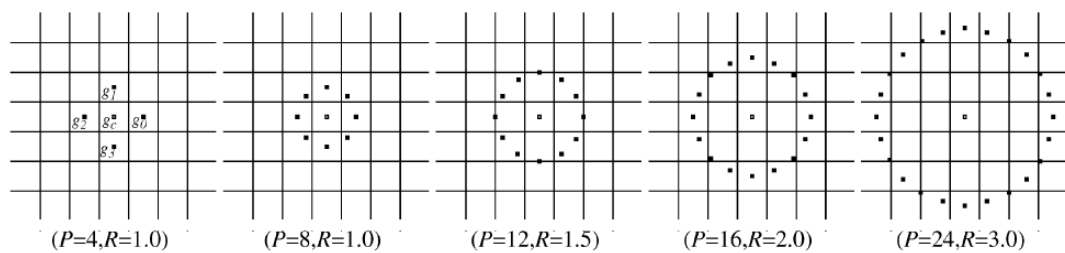
Figure 3.9: An illustration of multi-scale local binary pattern of with various sampling diameters and radiuses, where P is the number of sampling points, and R is the radius of sampling circle [132].

A multi-scale local binary pattern is obtained by concatenating the occurrence

histogram of local binary patterns extracted at different scales. $LBP_{P,R}$ represents a local binary pattern operator with P sampling points and sampling radius R. For example, the $LBP_{4,1+8,1}$ operator constructed from $LBP_{4,1}$ and $LBP_{8,1}$ has 272 bins. The $LBP_{4,1}$ operator generates an occurrence histogram of $2^4 = 16$ bins, and the $LBP_{8,1}$ operator generates a histogram of 256 bins. The final occurrence histogram of $LBP_{4,1+8,1}$ concatenates the occurrence histogram generated by both operators, as demonstrated in Figure 3.10.
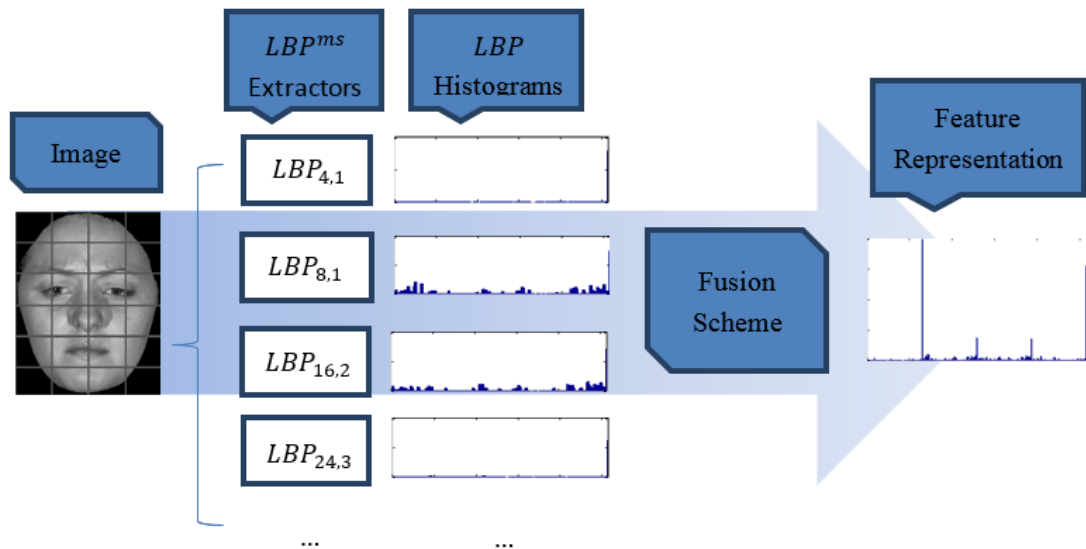
Figure 3.10: Feature generation using multi-scale local binary pattern operator. $LBP_{P,R}$ represents a local binary pattern operator with P sampling points and sampling radius R.

3.5.2 Local ternary pattern and its variants

3.5.2.1 Local ternary pattern

The local ternary pattern-based feature extractor operates on a 3×3 local neighbourhood. Instead of binarizing the local neighbourhood as the local binary pattern operator does [155], the local ternary pattern operator introduces a threshold t

which is used to remove the noise in a local patch of image and reduce the amount of noise contributing to the generated local ternary pattern. The local neighbourhood is thresholding into three ranges.

The thresholding function can be generalized as in Equation 3.7:

$$G(c, p_i, t) = \begin{cases} 1, & c \geq p_i + t \\ 0, & |c - p_i| < t \\ -1, & c \leq p_i - t \end{cases} \quad 3.7$$

Originally, the local ternary pattern should consist of $3^8 = 6561$ bins. Instead of using a three-value coding technique, Tan [154] separates the original ternary pattern into upper band and lower band, and then applies the same coding mechanism as the local binary pattern operator, as was explained in Section 3.6.1.1.

Mathematically, the upper pattern can be formulated as in Equation 3.8.

$$G(c, p_i, t) = \begin{cases} 1, & c \geq p_i + t \\ 0, & c < p_i + t \end{cases} \quad 3.8$$

The lower pattern can be formulated as in Equation 3.9.

$$G(c, p_i, t) = \begin{cases} 1, & c \leq p_i - t \\ 0, & c > p_i - t \end{cases} \quad 3.9$$

Figure 3.11 illustrates how a local ternary pattern code is divided into upper band and lower band. As a result of the partition, the local ternary pattern operator generates an occurrence histogram of $2^8 \times 2 = 512$ bins.
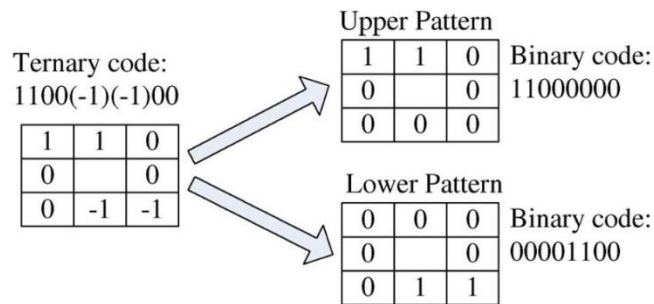
Figure 3.11: An illustration of the coding scheme adopted for coding the upper and lower patterns of a local ternary pattern.

3.5.2.2 Uniform local ternary pattern

The measure of 'uniformity' was first introduced in evaluating and reducing local binary patterns. Since a similar coding scheme is deployed in the generation of the binary code of a local ternary pattern, the same scheme is adopted for reducing the dimensionality of the local ternary pattern [132], [154]. By restricting the 'uniformity' of corresponding upper and lower patterns of the local ternary pattern to a maximum of 2 respectively, the total number of bins for the resulting occurrence histogram of the local ternary pattern is reduced to $59 \times 2 = 118$. A detailed description of uniform patterns can be found in Section 3.6.1.2.

To obtain the uniform local ternary pattern the following procedure, (illustrated in Figure 3.12) is executed:

1. Local ternary pattern operator is applied on the image patch.

2. Extracted local ternary pattern is divided into upper and lower patterns.

3. 'Uniformity' measure and uniform mapping is applied on upper and lower patterns respectively, which assign each uniform pattern with new bin labels, and sort non-uniform local ternary patterns into the same miscellaneous bin.

4. The reduced upper and lower pattern occurrence histograms are concatenated to form a uniform local ternary pattern. Consequently, an occurrence histogram with 118 bins is generated.
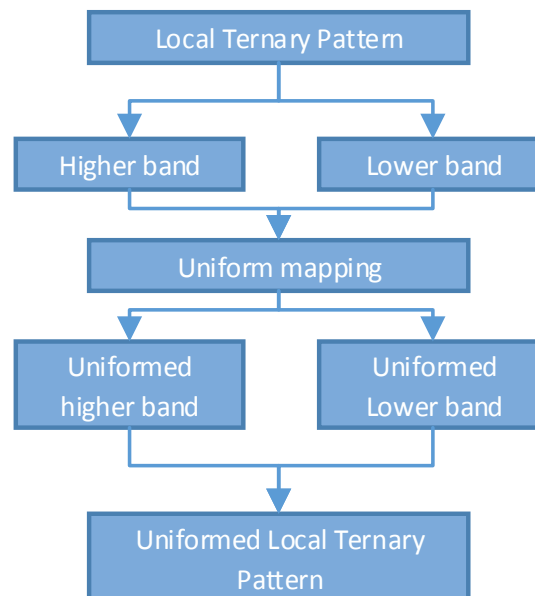
Figure 3.12: The generation of the uniform local ternary pattern.

3.5.2.3 Multi-scale uniform local ternary pattern

Motivated by the idea behind multi-scale local binary patterns, the local ternary pattern operator is extended to extract local ternary patterns at multiple scales via selecting sampling points in a circular manner with a specified radius. Equations 3.5 and 3.6 are applied to calculate the coordinates of sampling points on the sampling circle. In addition, the uniform mapping is applied on local ternary patterns extracted at each scale to reduce dimensionality of the final multi-scale local ternary pattern.
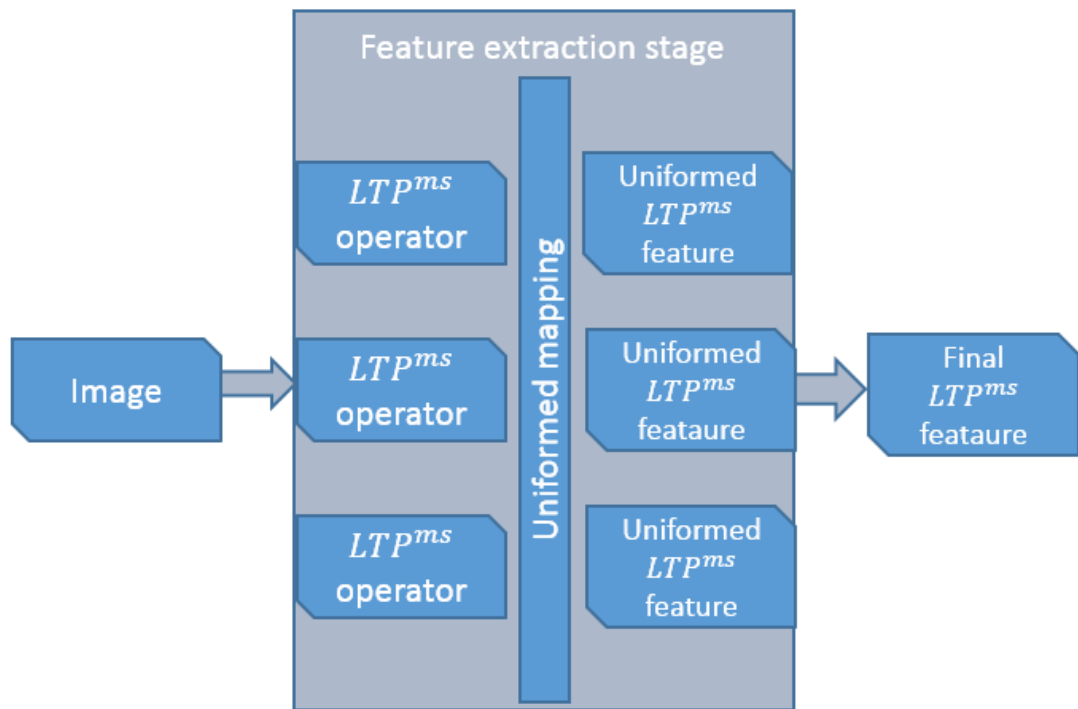
Figure 3.13: Feature extraction procedure of multi-scale local ternary pattern operator.

As illustrated in Figure 3.13, the following steps are executed to extract a multi-scale local ternary pattern:

1. Local ternary pattern operators of different sampling radius are applied on the image patch to extract multiple local ternary patterns.

2. Once multiple local ternary patterns are generated, the uniform mapping is applied respectively on each local ternary pattern.

3. Finally, the obtained uniform local ternary patterns are concatenated together to form a resulting multi-scale local ternary pattern.

3.6 Experimental setup and results analysis

In order to explore the usability and robustness of our proposed universal multi-view facial expression recognition system further, several experiments have been carried out. These, and an analysis of results, are presented in this section.

3.6.1 Data preparation

The *BU-3DFE* database contains 2,500 3D geometric shapes of the expressive face performed by the participants and 2,500 texture images of the expressive face, which were captured simultaneously. With the aid of the Simulink 3D Animation toolbox of Matlab [118], all six prototypic facial expressions are reconstructed from the provided facial texture images and 3D geometric shape models, including *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*. Through rendering and manipulating the 3D models (i.e. as described in Chapter 3), a facial expression dataset of six facial expressions with 4 intensity levels is created (where levels of 1 and 4 represent the least and the most intensive facial expression images respectively as described in Chapter 3). Facial expression images are simulated from 35 points of view within a pan angle range of $\pm45°$ and a tilt angle range of $\pm30°$ with an interval of $15°$, resulting in a total of 84,000 images. To normalize the size of the head in the resulting images, during the simulation process the distance between the rendering 3D shape and the camera is manually selected for each subject in the database via inspection of the model of the neutral facial expression, which means that an optimum parameter value is selected for each subject – a total of 100 values to cover 100 subjects, ensuring that a similar size of facial region in the simulated images at the frontal view is produced for all subjects. The distance parameter is used for all image generation processes of a particular subject regardless of changes of points of view. The 3D face model is rendered in the front of the camera, and therefore the head is always centred in the resulting images. The illumination is uniformly controlled across the entire rendering process. Finally, the simulated RGB images are converted to greyscale images, and

proportionally resized to 128*96. Figure 3.14 shows some examples of simulated facial expression images of one subject in the generated facial expression datasets.
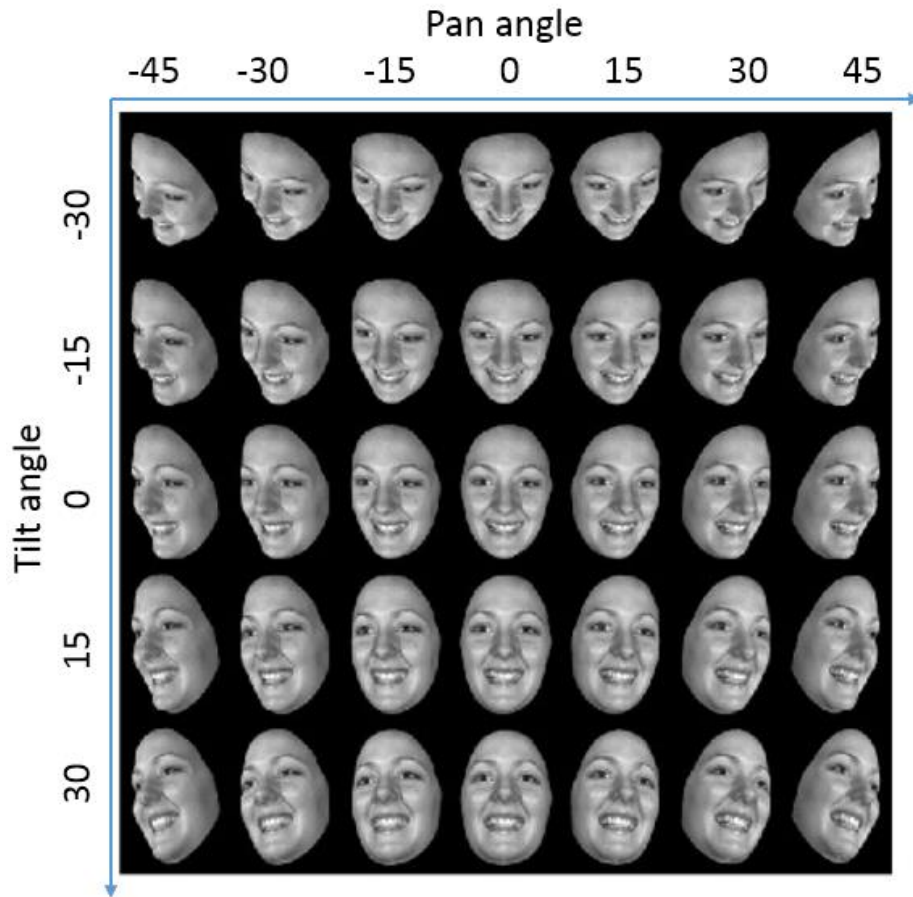
Figure 3.14: Examples of simulated facial expression images from one subject captured at 35 views at intensity level 4: 7 pan angles (-45°, -30°, -15°, 0°, 15°, 30°,45°), and 5 tilt angles (-30°, -15°, 0°, 15°, 30°) are used.

3.6.2 Pre-processing

Limited pre-processing is carried out to control the facial expression images generated by simulation, because the quality of 3D data contained in the *BU-3DFE* database is consistent and fairly controlled by the publisher, and the 3D image rendering process is also supervised. Consequently, only image normalization is carried out, using Equation 3.10:

$$f = \frac{255 \times floor(i(x,y) - L_{min})}{L_{max} - L_{min}} \qquad\qquad 3.10$$

$floor$ is an operator that rounds a number towards zero. L_{max} and L_{min} represent the maximum and minimum intensity levels in the image I. $i(x, y)$ is a pixel in the image with coordinates (x, y).

After normalization, the resulting image f will have an intensity range of $[0, 255]$.

3.6.3 Local ternary pattern and its variants

3.6.3.1 Holistic local ternary pattern

In the following experimental setup, a thorough investigation of holistic local ternary patterns with various specifications is conducted. We examine the local ternary pattern operators with 8 different extraction scales from radius 1 to 8 and 8 tolerance thresholds, with values of 2.5, 5, 7.5, 10, 15, 20, 50, and 100 which are tagged 1 to 8 respectively in all the graphs presented in this chapter. It should be noted that, as can be seen above, the tolerance thresholds, t, selected are not evenly spaced. In Figure 3.15, the overall classification accuracy of the system is demonstrated. The number of sampling points is fixed to 8 throughout this thesis because this setting allows the uniformity measure to be applied on the extracted local ternary patterns so that a more compact feature is available for further exploration in the later studies included in this thesis.
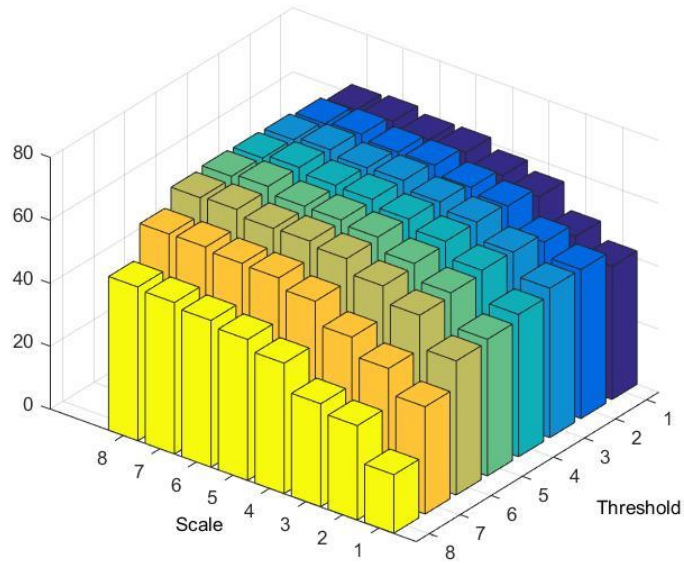
Figure 3.15: Overall performance in term of classification accuracy of local ternary pattern as a feature representation for universal multi-view facial expression recognition. The scales and threshold settings are explained in the main text.

	Average	Best
Accuracy	55.99%	66.81%

Table 3.1: Classification accuracy of holistic local ternary pattern representation

As shown in Table 3.1, the average performance of the basic local ternary pattern as a feature representation across 8 scales with 8 various threshold settings is 55.99%, and the best accuracy of 66.81% is achieved at scale 7 ($r = 7$) with tolerance threshold set to 15 ($t = 15$) . It is apparent from Figure 3.15 that the performance of local ternary pattern for universal facial expression recognition task has been dramatically influenced by both the scale of the local ternary pattern operator and the tolerance threshold selected.
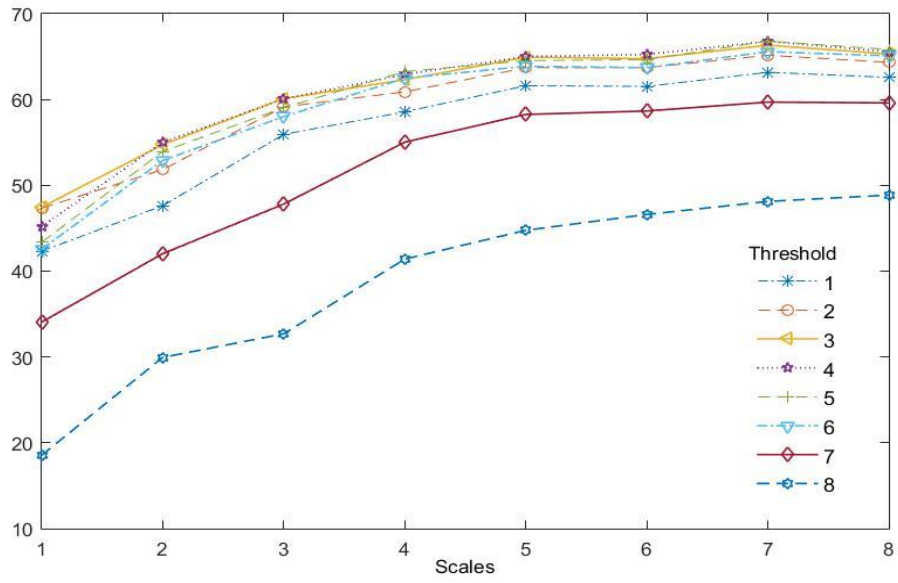
(a)
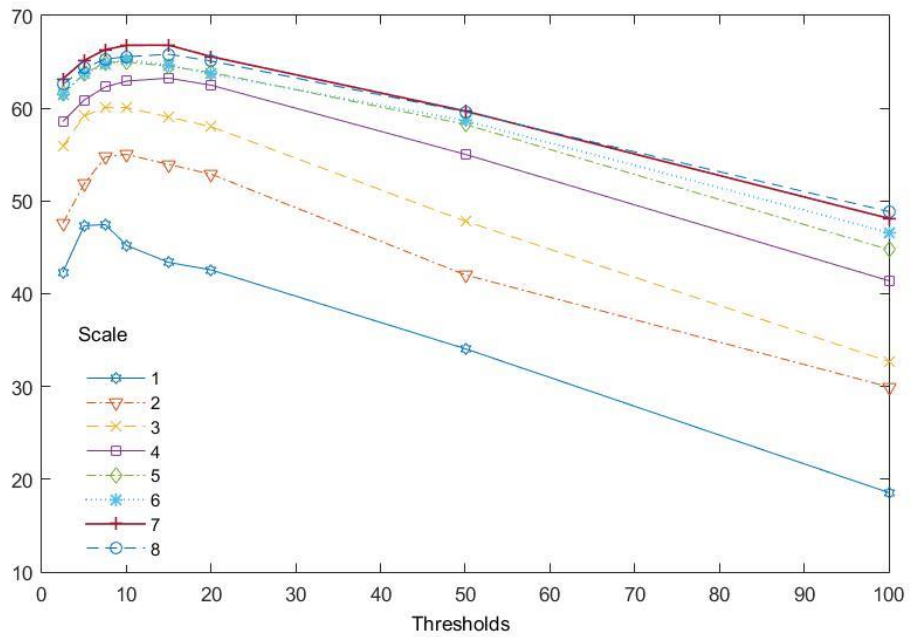
(b)

Figure 3.16: The performance curve of the universal facial expression recognition system: (a) w.r.t scales of the local ternary pattern operator; (b) w.r.t. the tolerance thresholds t selected for the local ternary pattern operator.

Graph (a) in Figure 3.16 illustrates how the proposed universal facial expression recognition system performs with respect to 8 sampling radius. It is strongly indicated in the performance curves that the LTP features extracted at scale 1 generate the worst performance. As the sampling radius of the local ternary pattern increases, the overall performance of the system also increases regardless of which tolerance threshold is selected. The best system performance is observed at scale 7 for all tolerance threshold settings, except tolerance threshold 1, for which the best system performance is observed at scale 8. Moreover, for all tolerance thresholds except the tolerance threshold tag of 1 and 2, a large performance gain is observed in between sampling radius of 1 and 5. Further increments of the feature extraction scale exert a smaller change in the system performance, although the peak of the system performance is observed at scale 7, except for tolerance threshold 1 at scale 8.

Graph (b) in Figure 3.16 shows the performance curves of our universal facial expression recognition system with respect to the changes of tolerance threshold t. 8 tolerance thresholds are selected in total with 4 evenly spaced sampling points smaller than 10, and 2 sampling points in between 10 and 20. With these sampling setups, a near bell-shape performance curve is observed with a sharp upward slope on the left climbing interval and a more downward slope on the right. For all scales, the performance curve peaks between tolerance threshold of 7.5 and 10, except scale 1 between 5 and 7.5. After tolerance threshold settings of 15, the system performance starts to fall gradually for all scales.

In addition, an investigation of the difference of performance between the high band and the low band of local ternary pattern, which is termed as LTP_{low} and LTP_{high}, is presented as demonstrated in Figure 3.17.
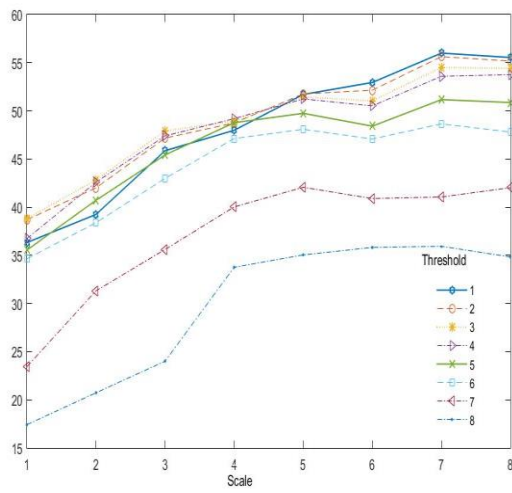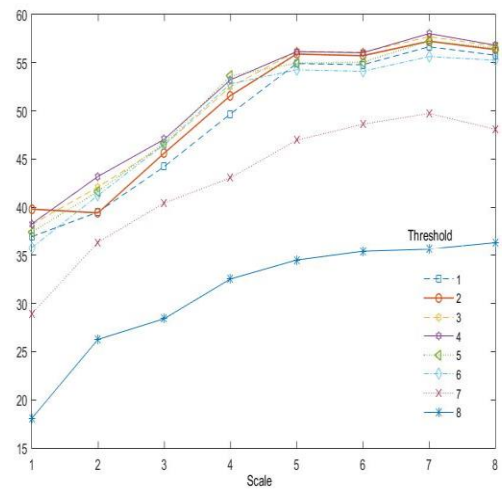
(a)   (b)
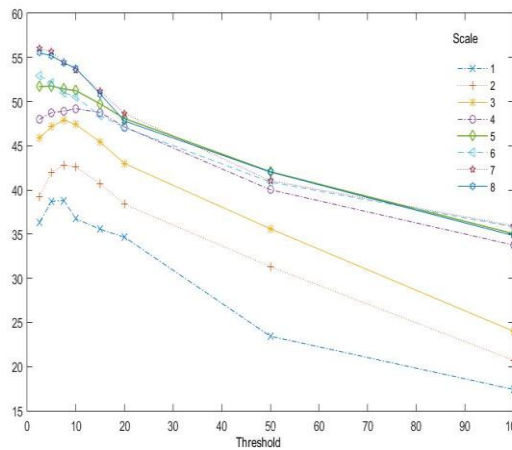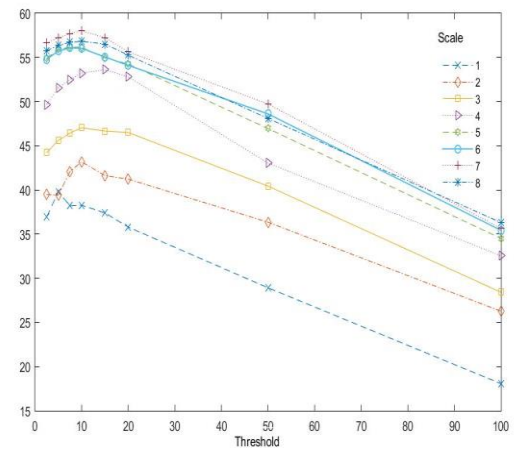
(c)   (d)

Figure 3.17: (a) & (c) The performance curve of LTP_{high} w.r.t scale and thresholds; (b) & (d) the performance curve of LTP_{low} w.r.t scale and thresholds.

By studying Figure 3.17, it is observed that the general trend of performance of the high and low bands of local ternary pattern is similar. More specifically, with the increment of the sampling radius, the performance for both bands of local ternary pattern increases and peaks at sampling radius of 7, and the system performance starts to decrease for thresholds larger than 20. Furthermore, it is observed that the

performance of LTP_{low} increases more sharply and the average performance of LTP_{low} is higher than LTP_{high} by 3%, at 49.81%.

3.6.3.2 Block based uniform local ternary pattern

Uniform local ternary pattern is a reduced local ternary pattern with 58 uniform pattern bins and 1 bin for all non-uniform patterns. In this experiment, a block based feature extraction of the uniform local ternary pattern is performed, which is labelled as $BBLTP$. The detailed description of feature extraction procedures can be found in Section 3.4.
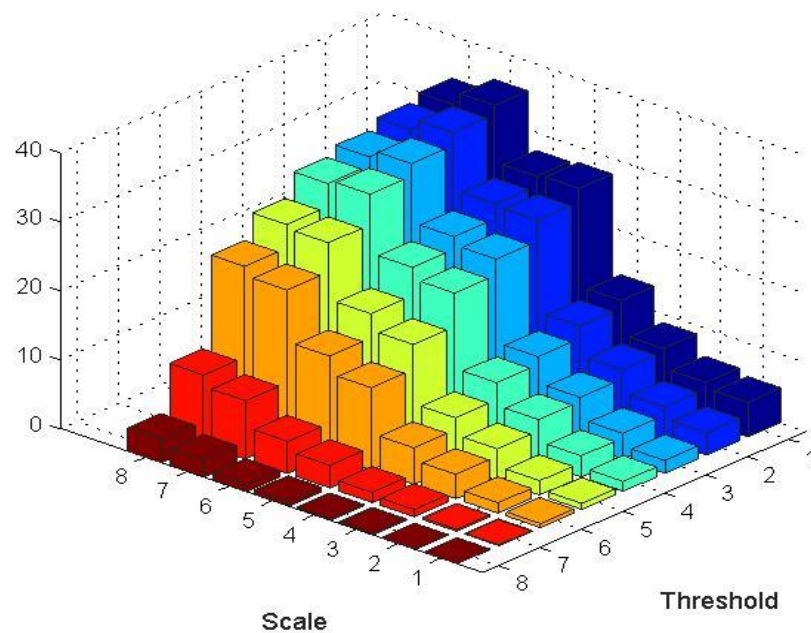
Figure 3.18: An inspection of the information loss in the conversion of basic local ternary patterns to uniform local ternary patterns.

Because the uniform mapping is employed to reduce the local ternary pattern, an experiment is carried out to measure the information loss during this process. The result of this uniformity evaluation is demonstrated in Figure 3.18. The percentage of

non-uniform local ternary pattern that is contained in the extracted *BBLTP* for all combinations of scale and threshold settings is measured. As shown in Figure 3.18, a lower percentage indicates that the non-uniform patterns comprise less in the total number of local ternary patterns. It is strongly indicated that, by modifying the parameters (i.e. the setting of scales and thresholds) the proportion of uniform local ternary pattern can be controlled. With a setting of low scale (i.e. a small sampling radius) and high threshold, the uniform patterns can comprise over 99% of all local ternary patterns extracted in the multi-view facial expression images, and vice versa, the uniform patterns comprise about 66% of all local ternary pattern at maximum.
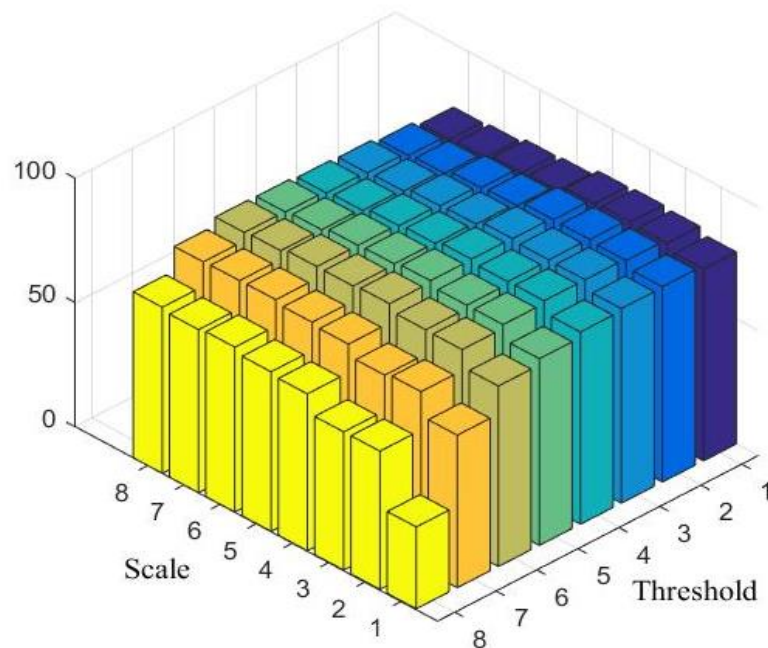
Figure 3.19: The overall classification accuracy of the block based uniform local ternary pattern w.r.t. various scales and threshold specifications.

As shown in Figure 3.19, the overall classification accuracy of *BBLTP* has improved significantly when compared with the holistic local binary pattern. By studying Figures 3.15 and 3.19, it is revealed that with respect to both operator settings of the *BBLTP* overall performance changes in a similar fashion as the holistic local ternary

pattern based system, in that the classification accuracy improves as the sampling radius increases when the tolerance threshold is at a reasonable small value. It shown in Figure 3.19 that, as the sampling radius increases, the overall classification accuracy increases and peaks at the scale of 4. In addition, minor changes of system performance are observed by continuing to increase the sampling radius of the block based local ternary pattern operator. With respect to threshold, the classification accuracy declines as the threshold increases in general. The system performance peaks at a threshold setting of 5, and a minor increase is observed between threshold setting of 1 to 5. After threshold setting 6 (i.e. $t = 20$), the classification accuracy starts to decrease. The average classification accuracy for the $BBLTP$ based universal multi-view facial expression recognition system is 76.31%, and the best performance is 83.23% obtained using the operator setting of 4 for the scale and threshold, as shown in Table 3.2.

$BBLTP$	Average	Best
Accuracy	76.31%	83.23%

Table 3.2: The average and best classification of block based uniform local ternary pattern representations.

By studying Figures 3.18 and 3.19, it can be seen that both non-uniform and uniform local ternary patterns contain descriptive information for the proposed classification task. It can be concluded that a block-based local ternary pattern with 99.99% of uniform patterns delivers classification accuracy of 67.1% and with 63.89% of uniform patterns gives a result of 75.35%. The best performance is observed with 90.07% of uniform patterns in the feature representation consisting of block based local ternary patterns. The above observation implies that to achieve the best performance of $BBLTP$, the balance of uniform and non-uniform patterns needs to be selected by controlling the threshold and scale of the operator.

Figure 3.20: Confusion matrix of classification results obtained with scale 4 and threshold setting of 4. Each figure is the classification accuracy (%).

Figure 3.20 demonstrates the classification confusion matrix among 6 expression classes. It is apparent that *fear* and *sad* are more difficult for the system to accurately classify, with a classification accuracy of 75.67% and 77.98% respectively, while *surprise* is the most distinguishable class with an accuracy of 93.19%. The facial expression of *fear* is misclassified with all other expression to a similar extent. *Anger* is most misclassified with *sad*, and *happy* is misclassified with *fear* the most. *Disgust* is misclassified with *anger* and *fear* to the same extent.
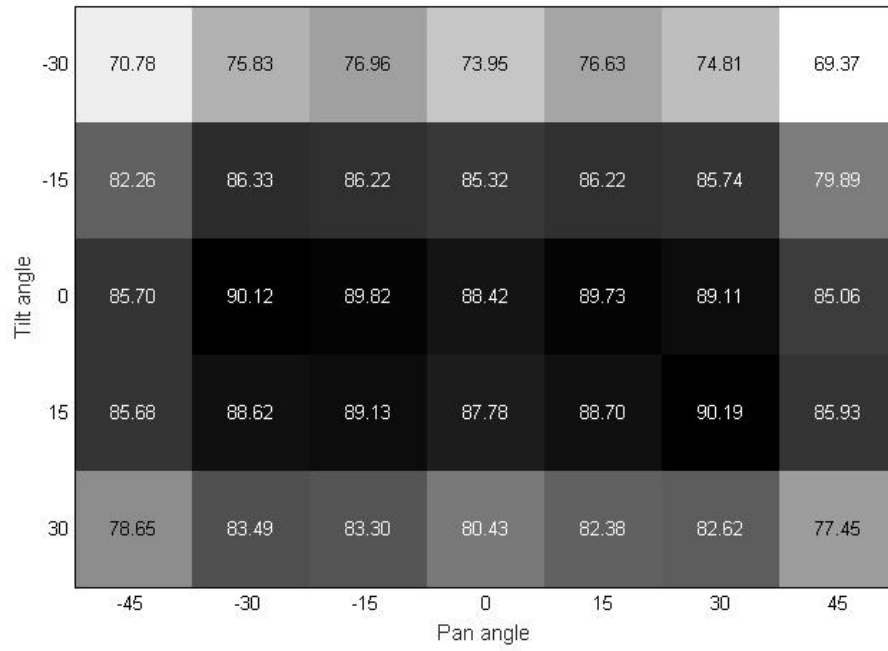
(a)
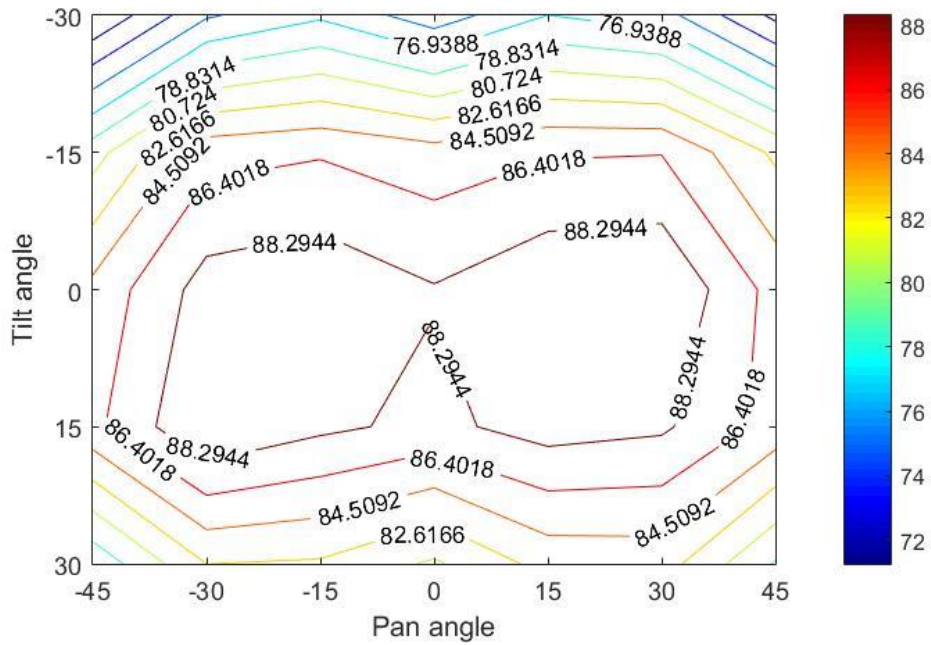
(b)

Figure 3.21: (a) The classification accuracy at each view in percentage; (b) the contour map, which demonstrate the change trend w.r.t. both tilt and pan angles.

By studying Figure 3.21, it can be concluded that facial expression recognition is influenced by the tilt and pan angles, and the frontal view is not the optimal angle for facial expression recognition, which coincides with the conclusion made by other researchers [159]. The reason behind this phenomenon might be that, at the frontal view, some facial features are less visible. However, a small variation in head pose can makes some informative features at particular areas of face become visible to the system, such as the cheek and triangle area of face. In addition, it is demonstrated in Figure 3.21 (b) that the tilt angle exerts a larger influence on the classification accuracy compared with the pan angle. A camera view from a point above the subject's face affects the performance of the system the most, and all the worst performance figures are observed at these viewpoints. In light of the trend of classification accuracy indicated in the contour map of Figure 3.21 (b), the pan angle's influence over the performance of the system can be observed as near symmetrical, and the classification accuracy peaks at $\pm 30°$, and starts to drop for pan views larger than $\pm 30°$.
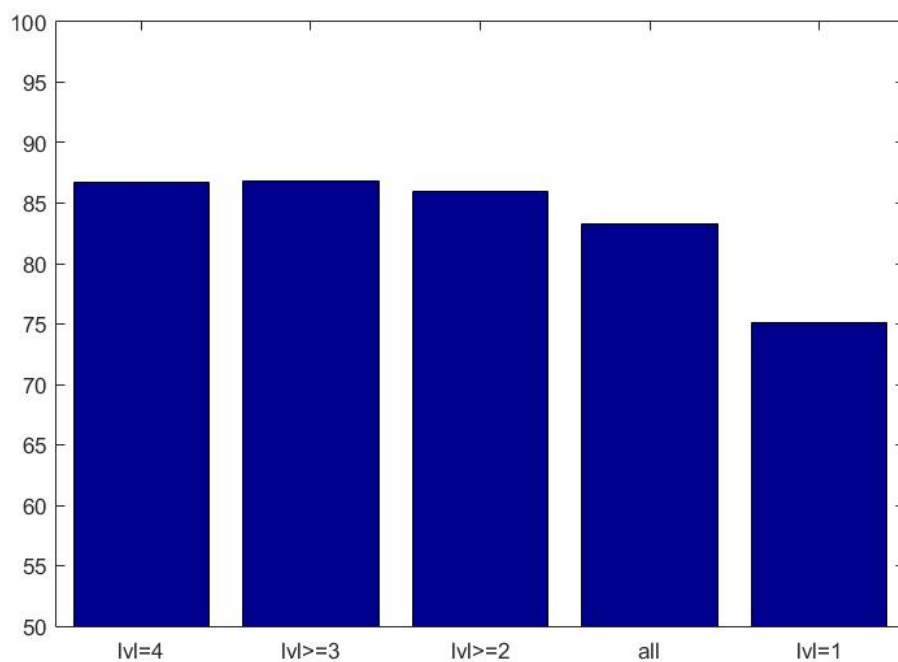
Figure 3.22: The classification accuracy of the proposed facial expression recognition system w.r.t. the changes of intensity levels. *lvl* designates the intensity level

In Figure 3.22, it is observed that the overall classification accuracy drops slightly as the total number of intensity levels of facial expressions the proposed system can handle increases. The classification accuracy observed at the lowest intensity level is 75.15%, which is lower than the accuracy obtained at the highest intensity by 11.5%. As shown, the classification accuracy does not decrease sharply (the accuracy does not drop below 85%) until the recognition at intensity level of 1 is included. Thus, it can be concluded that the overall system performance is significantly affected by the recognition accuracy obtained at lowest intensity.

3.6.3.3 Multi-scale local ternary pattern

According to the findings revealed in the block-based local ternary pattern classification experiment, the threshold setting of 4 is utilized for extracting the multi-scale block based local ternary pattern ($BBLTP^{ms}$) for universal multi-view facial expression recognition. The $BBLTP^{ms}$ consists of 8 local ternary pattern operators with 8 different sampling radius, which are applied simultaneously to extract $BBLTP$ patterns. Then the F-score feature selection algorithm is applied to reduce the dimensionality and redundancy in the completed $BBLTP$ feature set. Finally, the selected features are concatenated to form the final representation for the expression.

	Accuracy
$BBLTP^{ms}$	82.49 %

Table 3.3: The overall performance of universal multi-view facial expression recognition using the $BBLTP^{ms}$ representation

As shown in Table 3.3, the overall performance of the $BBLTP^{ms}$ representation is 82.49%. Comparing this with the classification result obtained with single $BBLTP$

pattern, the $BBLTP^{ms}$ representation is 6% higher than the overall average performance of the $BBLTP$ operators. The classification accuracy of $BBLTP^{ms}$ is similar to the $BBLTP$ operators with the best setting in terms of classification accuracy.

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	79.61	4.14	2.94	1.27	10.71	1.31
Disgust	6.14	81.39	4.29	2.46	3.38	2.35
Fear	4.04	6.44	73.22	7.33	4.89	4.08
Happy	1.24	2.24	5.98	88.73	1.06	0.76
Sad	12.61	2.23	3.30	1.09	79.53	1.24
Surprise	1.09	1.65	2.60	0.88	1.33	92.46

Figure 3.23: The classification confusion matrix for the proposed system with $BBLTP^{ms}$ representation.

Figure 3.23 that *fear* is the most confused facial expression for the proposed system setup and *surprise* is the easiest facial expression to recognize for the proposed system. Among all facial expressions, *anger* and *sad* are the most confused facial expressions, while *fear* is confused with all other expressions to same extent. *Surprise* is confused with *fear* by 2.6%.
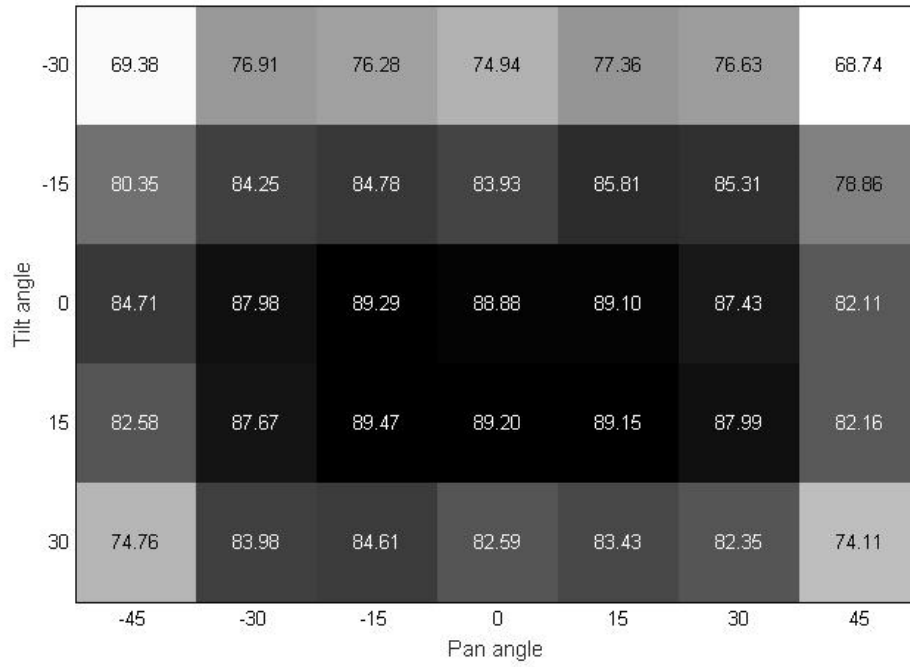
(a)
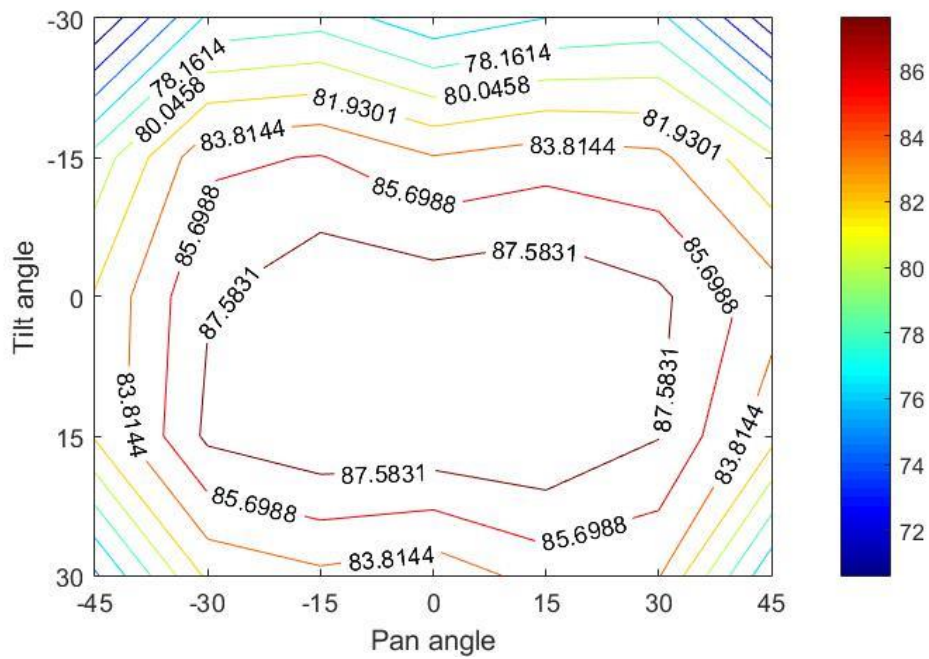
(b)

Figure 3.24: The classification accuracy of $BBLTP^{ms}$ feature representation. (a) The classification accuracy at each view in percentage; (b) the contour map, which demonstrates the change trend in w.r.t. both tilt and pan angles.

As demonstrated in Figure 3.24, the classification accuracy of the proposed system setup is affected by the viewpoint of the camera, which is, in other words, equivalent to head pose in a facial image with a completely opposite pan angle (i.e. as in a mirror image). From Figure 3.24 (a), it can be concluded that the tilt angle's impact on the overall system performance is larger than the pan angle variations, which coincides with findings summarized in view analysis of the block based local ternary pattern with scale and threshold parameter setting of 4 and 4, respectively. It is also reflected in the contour map of the classification accuracy that the performance drop is larger with a "looking down" view than with a "looking-up" view. In addition, with this system setup, the best performance is observed at pan angle of $\pm 15°$, which is not consistent with our previous observation in Figure 3.21. In addition, it should be considered that the average difference in system performance obtained for the frontal view and $\pm 15°$ is about 1% with all tilt angles, which implies that recognition at optimal pan views of $\pm 15°$ will not considerably increase system performance. A combination of extreme negative tilt angle and pan angles significantly restricts the performance of the proposed system, and reduces the overall classification accuracy to 69.4% and 68.7% as shown in Figure 3.24.

According to the experimental results illustrated in Figure 3.25, the classification accuracy is significantly affected by the proposed system's performance at the intensity level of 1. The difference of classification accuracy achieved by the proposed system between level 4 and level 1 is 13.74%, which is even larger than the sole local ternary pattern with specific threshold and scale as shown in Figure 3.25. No significant difference of classification accuracy is observed when intensity level 1 is excluded from scope of classification and the overall classification accuracy of the proposed system can reach 86.42%, which is comparable with the performance using the $BBLTP$ feature representation, at 86.5%. By studying Figures 3.22 and 3.25, it can be concluded that the performance difference between $BBLTP^{ms}$ and $BBLTP$ is caused by the difference of classification accuracy obtained at intensity level of 1.
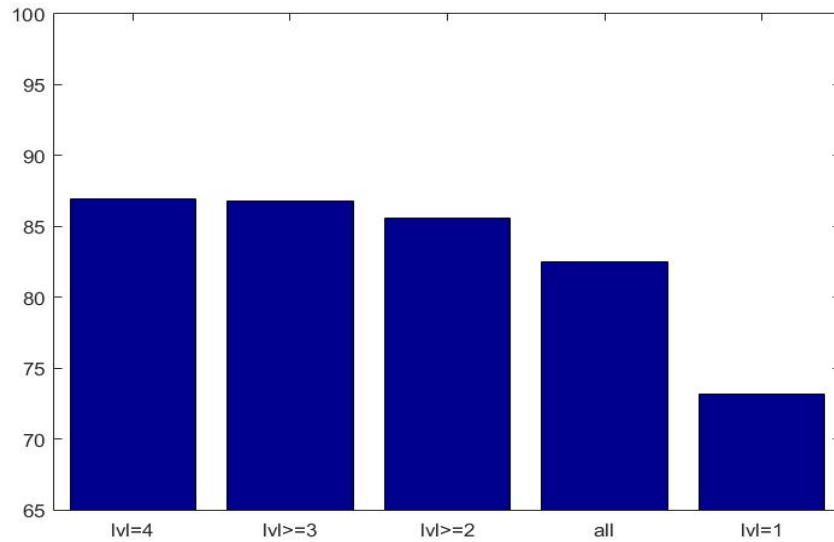
Figure 3.25: The classification accuracy of the proposed facial expression recognition system w.r.t. the changes of intensity levels (lvl) using $BBLTP^{ms}$.

3.6.4 Experiments on other facial expression databases

To further investigate and verify the validity and universality of our proposed approach for multi-view facial expression recognition, our approach is also tested on three other databases, including the Japanese female facial expression database (*JAFFE*) [69], the extended Cohn Kanade Database (*CK+*) [114], and our in-house multi-view facial expression database. The *JAFFE* database and *CK+* database are two widely available facial expression databases online and popular in reporting the general performance of a proposed facial expression recognition system in the facial expression recognition research community, thus testing our approach on these databases would make our method comparable with the state-of-the-art facial expression recognition systems in the literature and justify the robustness and universality of our proposed system. Our in-house multi-view facial expression database is a small database which is collected specifically for this project. Detailed information about these databases can be found in Chapter 3, where general information, data collection protocol, validity of the data, and availability of the databases will be explained.

3.6.4.1 General feature extraction procedures

The feature extraction process for the datasets selected from these databases is different from the approach employed for the *BU-3DFE* database [116], which does not contain a face detector as the facial region can be easily located since they are established and located in the center of the image. To prepare facial images for classification on *CK+*, *JAFFE*, and our in-house databases, firstly, the Viola Jones' face detector [32], which is explained in Chapter 3, is applied to identify the facial region in an image. The face detector can detect the face with about 90% accuracy, and the incurred mistakes during this process are manual corrected to assure the validity of derived facial expression dataset. After the facial region is properly segmented, a normalization process is applied using Equation 3.10, which normalizes the intensity of the image into the range of [0, 255]. Finally, the $BBLTP$ and $BBLTP^{ms}$ operators are applied to extract the corresponding feature representation for the image.

3.6.4.2 The Extended Cohn-Kanade database

The Extended Cohn-Kanade (*CK+*) [114] database contains 593 sequences of facial expression data of seven different facial expressions consisting of *anger*, *contempt*, *disgust fear*, *happy*, *sadness* and *surprise*. However, 309 sequences are chosen for this study, excluding video sequences of the expression of contempt, because our system is designed to classify prototypic facial expressions rather than action units. The final dataset consists of 618 facial images, which comprise the first and last frames of each sequence. The first frame represents the neutral expression while the last frame is the apex of a facial expression, which implies that the neutral expression comprises 50% of the derived dataset.

CK+	Accuracy	Average
BBLTP	99.66%	98.14%
BBLTP^{ms}	99.6%	

Table 3.4: The classification accuracy of our proposed system on the *CK+* database.

The performance of the proposed system on the derived *CK+* dataset is presented in Table 3.4 and Figure 3.26. As demonstrated in Table 3.4, our proposed system based on both $BBLTP$ and $BBLTP^{ms}$ can classify all six facial expressions including the neutral expression with a substantially high accuracy. By studying Figure 3.26, it can be observed that with the increment of the scale of the operator, the extracted $BBLTP$ features become more stable and efficient in describing the characteristics of facial expressions. In addition, the $BBLTP$ operator is stably accurate with respect to both scale and threshold changes until the tolerance threshold exceeds 20. The corresponding classification confusion matrices for both systems are shown in Figure 3.27 (a) and (b). As indicated in the experimental results presented in Figure 3.27 (a) and (b), for both $BBLTP$ and $BBLTP^{ms}$ based systems, *happy* and *surprise* are the simplest expressions to classify and *sad* is the worst. The confusion matrices of these two systems on the *CK+* database are similar with slight difference in classifying *anger* and *sad*.
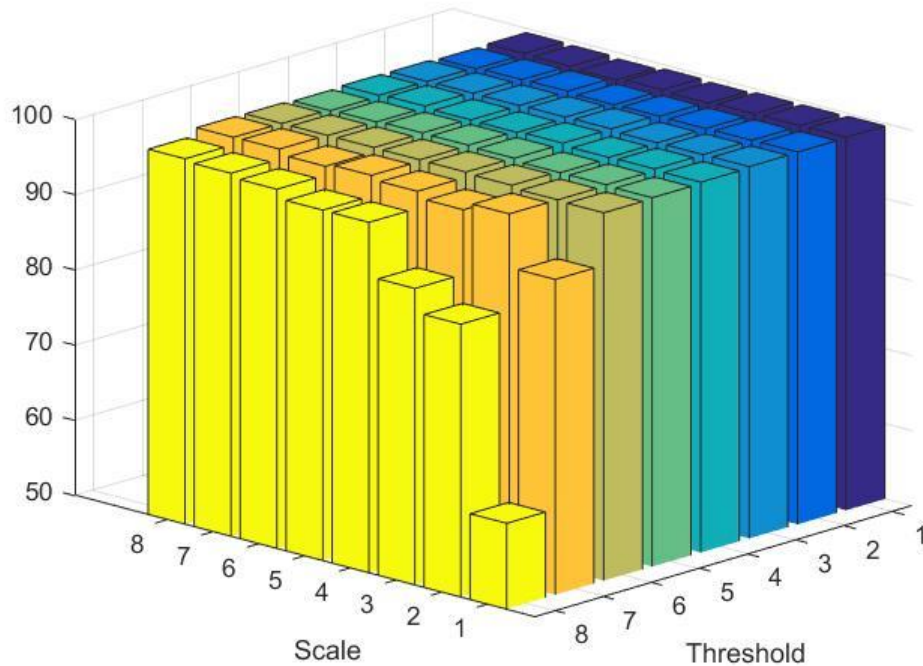
Figure 3.26: The performance accuracy of the proposed $BBLTP$ feature based system with respect to changes of scale and thresholds.

	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	98.18	0.00	0.00	0.00	1.82	0.00	0.00
Disgust	0.00	99.32	0.00	0.00	0.68	0.00	0.00
Fear	0.00	0.00	99.20	0.00	0.80	0.00	0.00
Happy	0.00	0.00	0.00	100.00	0.00	0.00	0.00
Neutral	0.00	0.00	0.00	0.00	99.97	0.00	0.03
Sad	0.00	0.00	0.36	0.00	1.79	97.86	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	0.00	100.00

(a)

	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	98.18	0.00	0.00	0.00	1.82	0.00	0.00
Disgust	0.00	99.32	0.00	0.00	0.68	0.00	0.00
Fear	0.00	0.00	99.20	0.00	0.80	0.00	0.00
Happy	0.00	0.00	0.00	100.00	0.00	0.00	0.00
Neutral	0.00	0.00	0.00	0.00	99.97	0.00	0.03
Sad	0.00	0.00	0.36	0.00	1.79	97.86	0.00
Surprise	0.00	0.00	0.00	0.00	0.00	0.00	100.00

(b)

Figure 3.27: (a) The confusion matrix of the $BBLTP$ based system; (b) the confusion matrix of $BBLTP^{ms}$ feature based system with tolerance threshold of 5.

3.6.4.3 The in-house multi-view facial expression database

To derive a testing dataset from our in-house multi-view facial expression database, both the first frame and peak frame from each valid video sequence is selected. Consequently, 392 facial expression images with 7 different pan views of 0, $\pm 30, \pm 60, \pm 90$ are used for this experiment, including six prototypic facial expressions and the expression of *neutral*. More detailed information about the in-house database is given in Chapter 3.

Our database	Accuracy	Average
BBLTP	95.641%	88.6899%
BBLTP^{ms}	94.87%	

Table 3.5: The classification accuracy of the proposed system on our in-house database.

According to the classification results obtained during the experiment as shown in Table 3.5, the *BBLTP* based system yielded an average classification accuracy of 88.69% across 64 different operator settings, and obtained the highest performance with the operator setting at scale of 5 and tolerance threshold of 10. The *BBLTP^{ms}* based system achieves a classification accuracy of 94.78%. This experimental result obtained on our in-house database has strongly implied that the proposed multi-view facial expression recognition system is capable of handling real data with reasonable performance with properly selected operator settings.
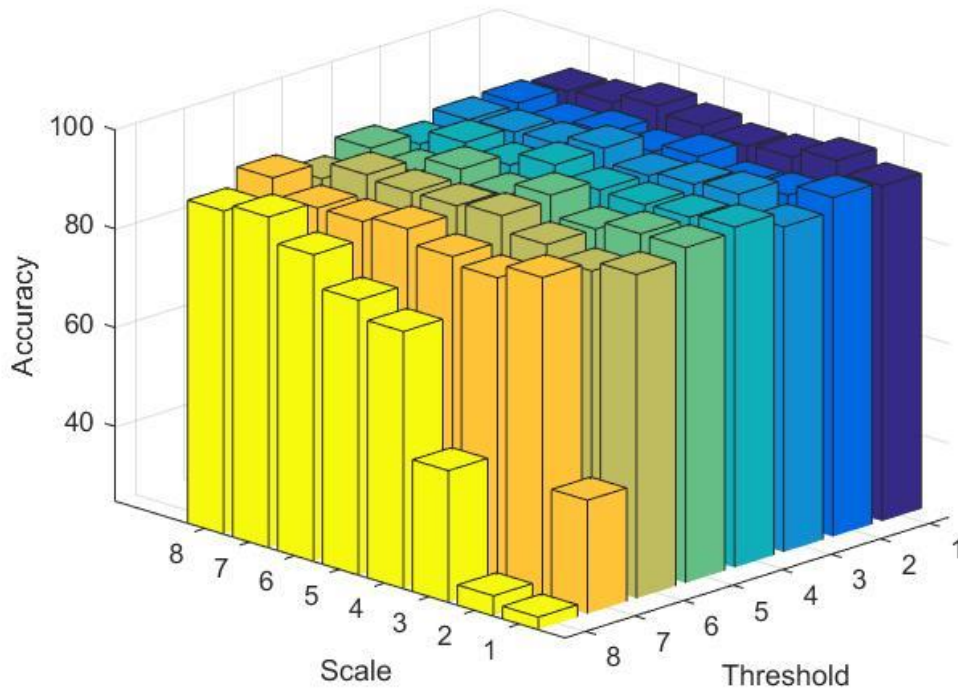
Figure 3.28: The classification accuracy trends of the proposed multi-view facial expression recognition system w.r.t. changes of the *BBLTP* operator settings.

As demonstrated in Figure 3.28 that the classification accuracy changes with respect to both changes of setting of scale and threshold. More specifically, when the tolerance threshold is not larger than 20, the proposed system can be reasonably stable regardless of the selection of scale. In addition, it is also observed that a larger scale setting of the operator generally delivers a better classification accuracy. A selection of operator setting of large threshold and small scale can result in an unstable or poor-performed system.

(a)

(b)

Figure 3.29: The classification confusion matrix is provided: (a) *BBLTP* based system; (b) *BBLTP^{ms}* based system.

Based on the confusion matrix illustrated in Figure 3.29 (a), it is shown that, for the *BBLTP* based multi-view facial expression recognition, the facial expression of *surprise* is the best recognized facial expression though all other facial expressions might be misclassified as *surprise* to a minor extent. The expression of *fear* is confused with *surprise*, which also ranked the second in terms of classification accuracy. Other facial expressions are classified with a similar accuracy. According to the confusion matrix shown in Figure 3.29 (b), the facial expression of *happy* is the best classified expression with a minor confusion with *neutral* while *anger* and *fear* are classified the worst. The expressions of *disgust*, *fear*, and *sad* have been misclassified as *anger* for around 5%.
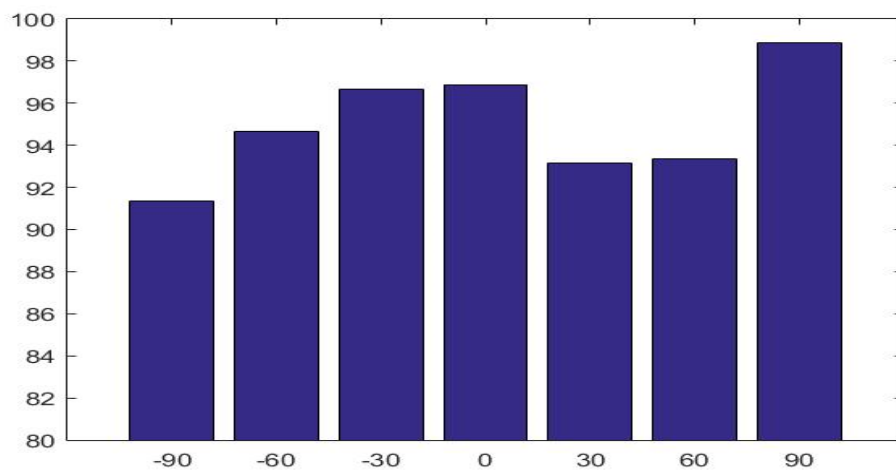
(a)
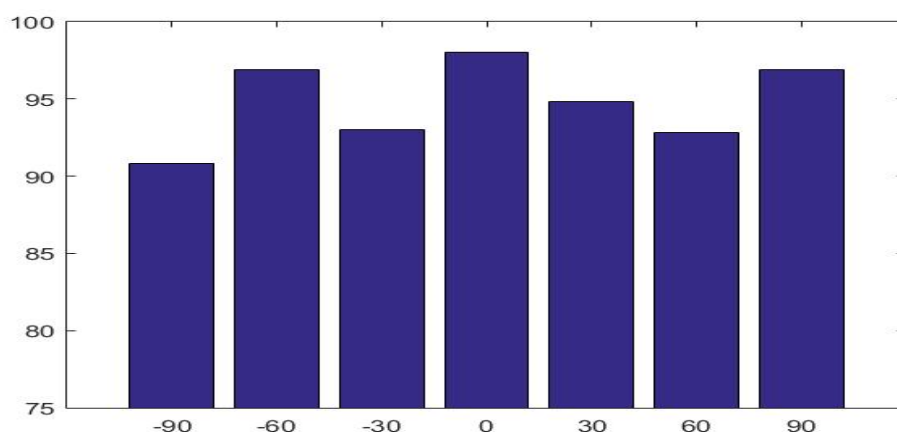
(b)

Figure 3.30: The classification accuracy of the proposed system with respect the changes of the pan views: (a) $BBLTP$ feature based; (b) $BBLTP^{ms}$ feature based.

Figure 3.30 (a) demonstrated the performance of the proposed *BBLTP* based system for various pan views. It is indicated by the experimental results that a pan view of 90 $^\circ$ is the best view for facial expression recognition for our system while the pan views of 0° and -30 $^\circ$ deliver a slightly worse performance. The proposed system is operating at over 90% accuracy across all pan views. Furthermore, as shown in Figure 3.30 (b), it can be concluded that with the increase of pan angle, the performance of the system degrades generally, and the frontal view (i.e. pan view of 0 $^\circ$) is the best view for facial expression recognition.

3.6.4.4 Japanese female facial expression database

The Japanese female facial expression (*JAFFE*) database contains 219 frontal facial expression images of six prototypic facial expressions and the neutral expression. In this experiment, the complete set of the *JAFFE* database is utilized. To extract the feature representation for the image, the general feature extraction procedures described in Section 3.7.4.1 is followed.

Our database	Accuracy	Average
BBLTP	99.62%	99.16%
BBLTP^{ms}	99.76%	

Table 3.6: The classification accuracy of $BBLTP$ and $BBLTP^{ms}$ feature based system on the *JAFFE* database.

As shown in Table 3.6, the performance of the *BBLTP* feature based system can classify all seven facial expressions at 99.16% accuracy and reaches the highest performance of 99.62% at scale 7 and tolerance threshold of 15. The $BBLTP^{ms}$ achieve a classification accuracy of 99.76%. In addition, it is shown in Figure 3.31

that with the increase of scale of the operator, the classification accuracy of the proposed *BBLTP* feature based system improves gradually in general while no obvious trend is observed with respect to changes of the threshold setting except when the tolerance threshold is set as 100, the system becomes unstable with respect to scales. The confusion matrices for the *BBLTP* and $BBLTP^{ms}$ feature based system are presented in Figure 3.32 (a) and (b) respectively. Based on the classification results shown in Figure 3.32 (a) *sad* is the worst classified expression as it is misclassified with three different expressions, while *fear* ranks the second worst with misclassification to *sad* and *surprise*. Regarding the $BBLTP^{ms}$ feature based system, *anger*, *happy*, and *neutral* continue as the best classified expressions by the proposed system. *Disgust* is misclassified with *anger* while *sad* and *surprise* is misclassified with *happy* as shown in Figure 3.32 (b).
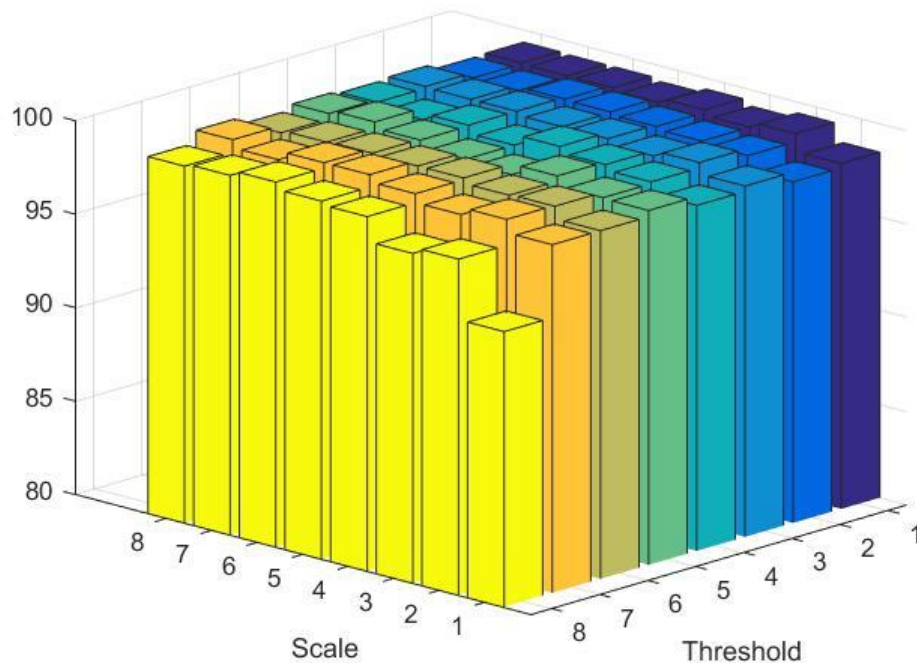
Figure 3.31: The classification accuracy the *BBLTP* feature based system yields w.r.t. changes of the setting of the *BBLTP* operator.

(a)

(b)

Figure 3.32: The confusion matrix for the proposed system using $BBLTP$ (a) and $BBLTP^{ms}$ (b) feature.

To sum up, according the experimental results and analysis presented, the primary findings observed can be summarized as follows:

1) According to the experimental results, our proposed system achieved a classification accuracy of 83.23% and 82.49% using the *BBLTP* and *BBLTP^{ms}* feature representation respectively, which outperforms the state-of-the-art multi-view facial expression recognition system in terms of accuracy by 5% across all intensity levels and by 10% excluding the lowest intensity level on an experiment operated on 84,000 images of 35 various views. The comparison of system performance of our proposed system and the state-of-the-art facial expression recognition system is provided in Table 3.9.

2) Another advantage of the proposed system is that it includes a noise tolerance threshold, which can be deployed to either control the uniformity ratio of the generated data and the illumination bias possess by a facial image. As a result, a more reliable feature representation can be devised.

3) *Fear* is the most difficult facial expression to be recognized for our proposed facial expression recognition system. Facial expressions of *anger* and sad are confused with each other using both *BBLTP* and *BBLTP^{ms}* feature representations.

4) Facial expression recognition at the lowest intensity level is the most difficult, and restricts the overall performance of the proposed system. Excluding the intensity level of 1 from the scope of the classification, our system can achieve a classification accuracy of over 86%.

5) For our proposed systems, the best view for facial expression recognition is observed at tilt angle of 0° and pan angle of −15°. Tilt angles exerts a more significant influence over the overall system performance compared with pan angle.

6) According to our preliminary experimental results obtained on *CK+*, *JAFFE*, and our in-house facial expression database, it can be concluded that the proposed novel multi-view facial expression recognition based on *BBLTP* and *BBLTP^{ms}* features can operate on real data with an outstanding performance with a proper selection of operator parameters. The proposed system has outperformed the state-of-the-art system by about 5% on both databases as shown in Table 3.7 and 3.8.

System	Zhang [112], facial movement feature	Huang et al. [160], component based feature descriptor	Wong [161], FEETS and HFTS	Ramirez Rivera et al. [162], LDN based approach	Our approach
Performance	94.48%	93.85%	95.87%	89.3%	99.66%

Table 3.7: Classification accuracy of the state-of-the-art facial expression recognition systems tested on *CK+* databases.

System	Zhang [112], facial movement feature	Shih et al. [110], DWT and 2D-*LDA*	Gu [163], Radial encoded Gabor jets	Lee et al. [164], sparse representation	Our approach
Performance	92.93%	94.13%	89.67%	94.7%	99.76%

Table 3.8: Classification accuracy of the state-of-the-art facial expression recognition systems tested on *JAFFE* databases.

FER System	Feature representation	Extraction method	Views	intensity	Total of images	Accuracy
Tang et al. [135]	*EHMM + SIFT*	Dense *SIFT*	7 pan angles and 5 tilt angels, 35 views in total	4	21,000	**75.3%**
Hu et al. [158]	Geometric shape coordinate displacement	Facial landmark based	5 pan angles (0°, 30°, 45°, 60°, 90°), 5 views	2, 3, 4	12,000	**66.5%**
Zheng et al. [133]	*SIFT* + 83 landmarks	Facial landmark based	5 pan angles (0°, 30°, 45°, 60°, 90°), 5 views	All intensity levels	12,000	**78.3%**
Moore and Bowden [136]	*LGBP + LBP*	Block based	5 pan angles (0°, 30°, 45°, 60°, 90°), 5 views	All intensity levels	48,000	**71.1%**
Usman Tariq et al. [134]	Dense *SIFT*+ *SSVQ + SQM* (max pooling)	Dense *SIFT*	7 pan angles and 5 tilt angels, 35 views in total	4	21,000	**76.34%**
Ours			7 pan angles and 5 tilt angels, 35 views in total	All intensity levels	84,000	**83.23%**
Ours	*BBLTP*	Block based		2, 3, 4	63,000	**86.5%**

Table 3.9: Classification accuracy of state-of-the-art multi-view facial expression recognition systems compared with our proposed facial expression recognition systems using *BU-3DFE* database.

3.7 Conclusion:

Following a thorough review of the foremost mechanisms for designing a multi-view facial expression recognition system and state-of-the-art multi-view facial expression recognition system, a new universal multi-view facial expression recognition system based on local ternary patterns is presented, and a comprehensive investigation is conducted to explore the influence of different specifications of the local ternary pattern on the overall performance of the proposed system in terms of achievable classification accuracy. Furthermore, the performance of the proposed system with respect to both pan and tilt angles and intensity levels is also explored.

In the following Chapter, a novel local descriptor, level of difference descriptor, is proposed to use as a supplement for the state-of-the-art local descriptor, which is weak in describing local appearance, along with a systematic exploration of the fusion of texture features.

Chapter 4

Fusion of local descriptors for universal multi-view facial expression recognition

In this chapter, the application of state-of-the-art texture descriptors, including local binary pattern, histogram of oriented gradient, and the gray level co-occurrence matrix, in devising a universal facial expression recognition system is examined. Then, an investigation of the robustness of the devised feature representation against variation of intensity levels and pan and tilt angles of the head is elaborated. Furthermore, a novel local descriptor is proposed, which designated as the level of difference descriptor, to use as a supplement to state-of-the-art local descriptors to further improve the performance of a system in terms of classification accuracy. Finally, the fusion of various texture features for devising a robust feature representation for universal multi-view facial expression recognition is presented.

Section 4.1 gives a general introduction to state-of-the-art texture features. In Section 4.2, the general framework of our proposed multi-view facial expression recognition system is presented. Section 4.3 thoroughly describes all the texture descriptors that are adopted in this study. In Section 4.4, the experimental results obtained from a series of experiments are systematically presented and analysed, and a brief summary of the experimental findings is also included. Section 4.5 briefly summarizes all the principal findings presented in this chapter together with some overall conclusions.

4.1 Introduction

As mentioned in Chapter 3, various state-of-the-art texture descriptors have been employed to derive a solution for the multi-view facial expression recognition problem, including local binary pattern operator [24], Gabor feature [130], scale invariant feature transform [131], histogram of oriented gradient [75], discrete cosine transform [76], speed up robust feature [37], gradient location and orientation histogram [74], and so on (examples of these texture descriptor based systems can be found in Chapter 3 Section 2). However, as demonstrated in Table 3.6 in Chapter 3, these individual texture descriptor-based 2D multi-view facial expression recognition systems have found classification with high accuracy of the six prototypic facial expressions to be challenging, not to mention the more complicated facial expressions, such as *anxiety*, *frustration*, and *depression*. Therefore, it is necessary to further improve the general performance of the multi-view facial expression recognition system. In this chapter, a comprehensive analysis of the state-of-the-art texture descriptors including histogram of the oriented gradient [75], gray level co-occurrence matrix [165], local binary pattern [24], and the fusion of these features in constructing a feature representation for a multi-view facial expression recognition system is examined. In addition, a novel local descriptor, which is called the level of difference descriptor, is introduced to use with the other state-of-the-art texture descriptors. The next section will explain the general framework of our proposed multi-view facial expression recognition system for the series of experiment described in this chapter.

4.2 Framework for the multi-view facial expression recognition system

In the light of the performance that the previous $BBLTP$ and $BBLTP^{ms}$ based multi-view facial expression recognition has achieved, as described in Chapter 3, the same system structure is employed for evaluation of the individual texture descriptor-based systems in order to examine of the performance of these state-of-the-art descriptor in

application of multi-view facial expression recognition. The detailed information about the structure of the proposed system can be found in Section 3.3 in Chapter 3.
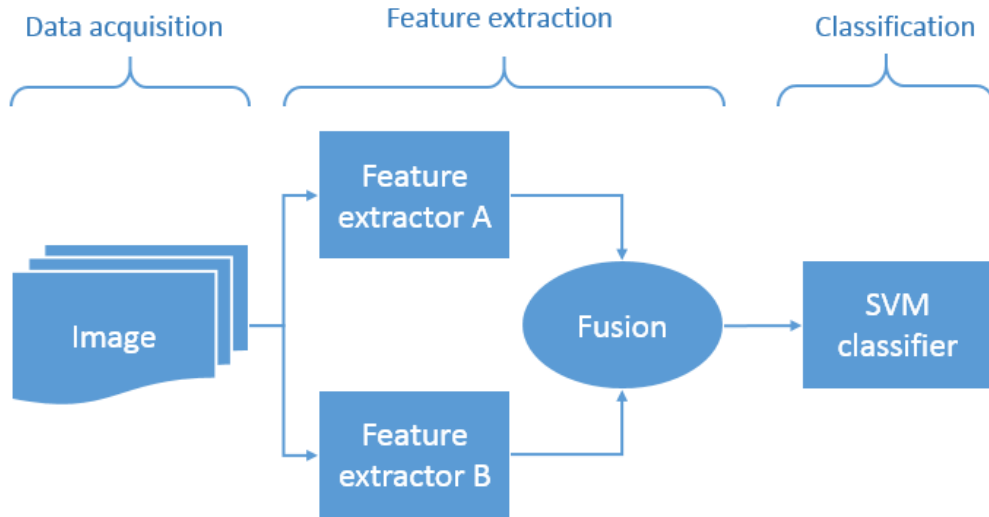
Figure 4.1: An illustration of the structure of the proposed universal multi-view facial expression recognition system.

For the exploration of the fusion of texture descriptor-based system, the data acquisition stage and classification stage of the proposed system remain the same as the system structure described in Section 3.3. With regard to the feature extraction stage, minor adjustments and an addition of the feature selection module is added to the previous proposed system, as illustrated in Figure 4.1. Two texture descriptors are deployed to extract a feature representation because the objective of this study is to find real time solutions for multi-view facial expression recognition in practice, which has set a limit of 0.04 second in feature extraction time for one image. After that, the two extracted texture feature representations are combined using the F-score feature selection algorithm, which is described in Section 3.5, in Chapter 3 with a proportional

fusion scheme, which is described in Section 4.4.7. The details of the employed texture feature are explained in Section 4.3.

4.3 Texture features

In this study, a number of texture descriptors are investigated, including local binary patterns and its variants, and local ternary pattern and its variants, level of intensity difference, and histogram of orientations, in the application of universal facial expression recognition. The details of the adopted texture feature descriptors are presented in this section, except for the local binary pattern, which has been thoroughly described in Section 3.6 in Chapter 3.

4.3.1 Level of difference descriptor

After experimenting with the state-of-the-art texture descriptors, such as the local binary patterns, local ternary pattern, it is observed that certain local descriptors adopt a process that compresses or discards the local appearance of the imagery region, and as a result, a significant amount of informative content in the local neighbourhood is omitted in the extraction scheme. It is therefore reasonable to argue that this appearance information might be used, together with the aforementioned texture descriptors, to improve the overall system performance. Thus, a new *level of difference* (*LOD*) descriptor is introduced, designed as a supplement for established local descriptors that are less informative in describing local pixel intensity.

Fundamentally, the level of difference descriptor is designed to statistically describe the difference of intensity in the defined local neighbourhood in general. First, an image region is divided into blocks, and then by calculating the difference of the intensity of each pixel with the average intensity of the entire block, each pixel casts its vote into a histogram of 16 bins corresponding to values ranging from -255 to +255,

which have an evenly spaced interval of 32. Specifically, the *LOD* pattern is extracted in the following steps, which is also briefly illustrated in Figure 4.2:

1) The average intensity of the block is calculated using equation 4.1:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad\qquad 4.1$$

In the above expression, n represents the total number of pixels in the selected sub blocks; X_i is a vector, which includes all the intensity values in the block.

2) All pixels are subtracted from the mean.

3) Each pixel casts a vote to an occurrence histogram, which consist of 16 bins and a bin width of 32. The resulting representation is a 1×16 feature vector for the block.
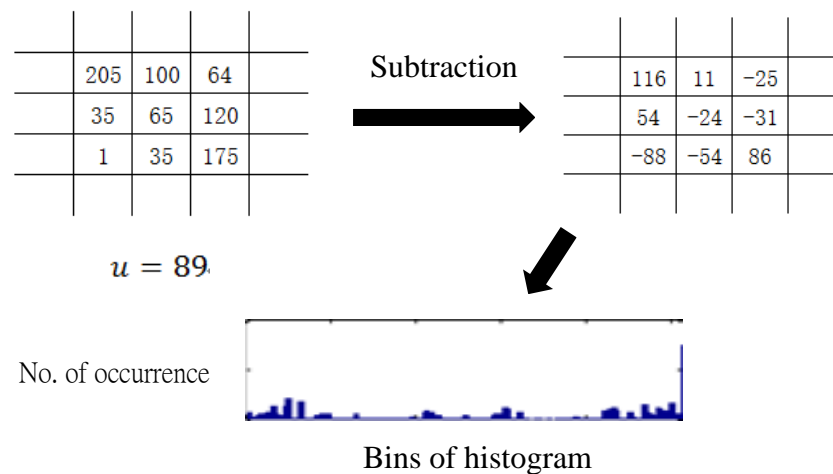
Figure 4.2: An illustration of feature extraction of the level of difference operating on a block size of 3. u is the average of all pixels' intensity values contained in the block.

4.3.2 Histogram of oriented gradient descriptor

The histogram of oriented gradient descriptor (*HOG*) is a local descriptor originally developed for detection of human by Dalal and Triggs [75]. The fundamental idea of the *HOG* descriptor is to collect and record the characteristics of the textural appearance and geometric shapes through the local intensity gradient while allowing a defined displacement tolerance of the pixels. More specifically, the *HOG* descriptor is a block-wise operator, which calculates the orientation of the gradient at each pixel in the local blocks, and accumulates the weighted orientation votes within the block into a histogram to form a texture representation. In our study, the feature extraction of the *HOG* feature is carried out in the following steps:

1) First, an image is densely extracted with an overlap of half block size.

2) Second, the feature extraction is carried out in a detection block. Each block is partitioned further into 4 cells of size of 8×8 pixels. The size of a detection block is 16×16 pixels.

3) For each cell, the local directional gradient is calculated, and then accumulated into a 1-dimensional histogram of oriented gradients, which separates the orientation range of $180°$ into 9 equally spaced bins.

4) All the histograms generated by cells within the block are then concatenated into a feature vector representing the block.

5) By concatenating all the densely extracted feature vectors in the image, the final feature representation is completed.

4.3.3 Gray level co-occurrence matrix and its statistics

The co-occurrence matrix (*GLCM*) and its statistics [165] is one of the earliest texture descriptors which still plays an important role in computer image analysis. It is an appearance-based operator with an assumption that pixel intensity level and the pairwise spatial relationship of pixels together possess a unique characteristic. The *GLCM* operator exhaustively examines an image region for all possible pair-wise intensity patterns of the same spatial relationship, and then counts the occurrence of each pattern in an x by x matrix, as illustrated in Figure 4.3, where x is the number of intensity levels [165]. One co-occurrence matrix, which is also referred to as the angular nearest neighbour gray tone dependence matrix, corresponds to a given spatial relation operator, which implies that one image patch could have multiple occurrence matrixes. To reduce the computational cost of calculating an image intensity spatial dependence matrix, generally a reduction of the range of intensity values is executed before calculating the matrix. Figure 4.3 shows an example of the calculation of the co-occurrence matrix.
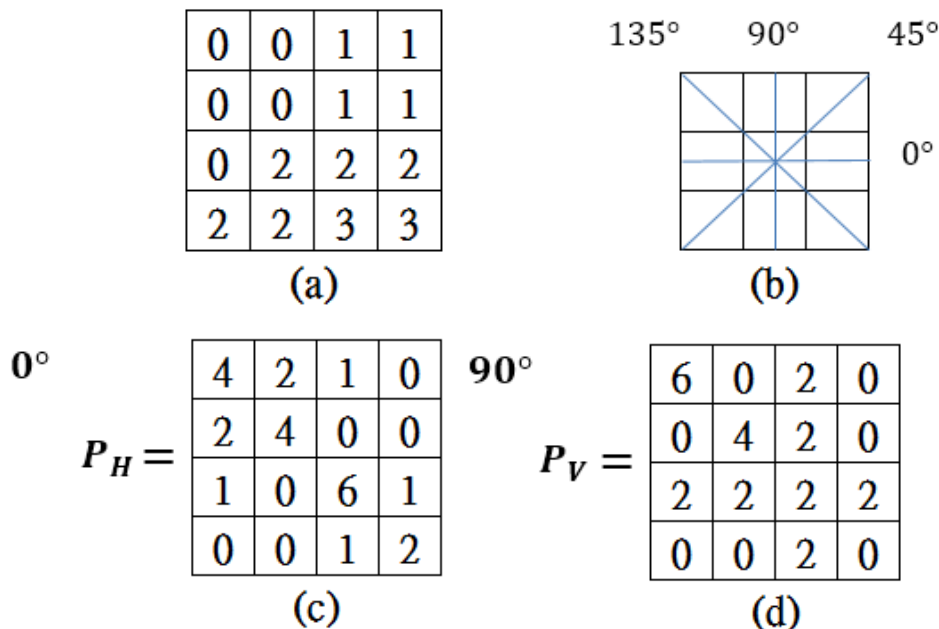
Figure 4.3: (a) A 4 × 4 image block with gray range of 0 to 3; (b) four directional relations; (c) P_H is the co-occurrence matrix calculated for 0° direction at distance of

1; (d) P_V is the co-occurrence matrix calculated for 90° at distance of 1 (taken from [165]).

4.4 Experimental setup and results analysis

4.4.1 Block based uniform local binary pattern
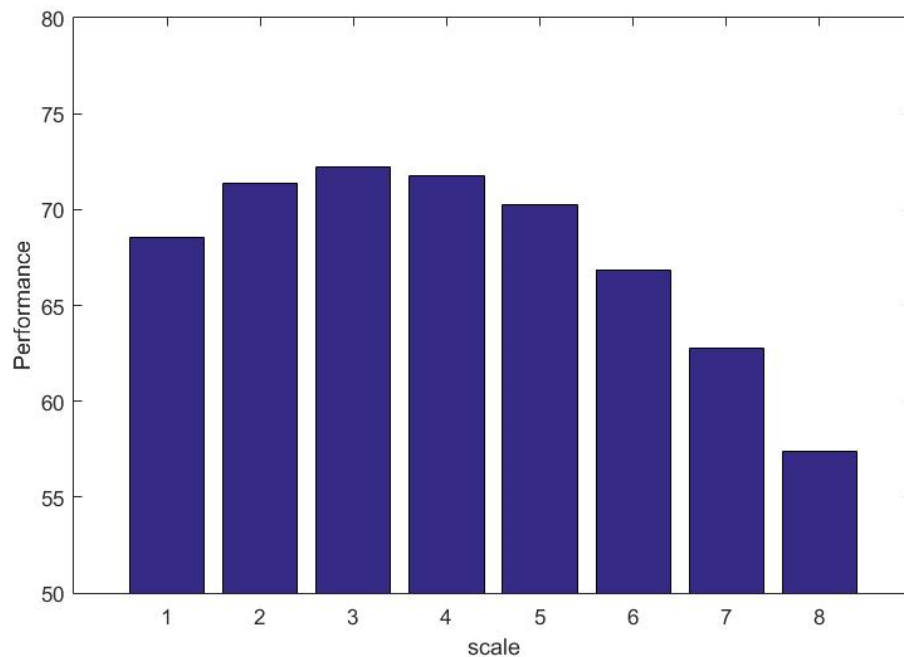
Figure 4.4: Classification accuracy (%) of the proposed *BBLBP* feature based system with local binary patterns extracted at 8 different scales.

A thorough investigation of the block based local binary pattern (*BBLBP*) for application in universal multi-view facial expression recognition is reported in this section. Eight different feature representations for the facial expression image

constructed using the local binary pattern extracted at 8 different scales is adopted in this experiment, and the performance of various configurations is illustrated in Figure 4.4. Distinct from the block based local ternary pattern whose performance increases as the scale of operator increases, this figure shows that the overall performance of the system based on *BBLBP* first increases, and peaks at the extraction scale of 3, and then starts to decrease as the scale of extraction of the local binary pattern operator continue to increase. To sum up, it is apparent from the Figure 4.4 that the scales of the local binary pattern operator has a significant influence on the performance of our proposed facial expression recognition system.

BBLBP	Average	Best
Accuracy	67.66%	72.25%

Table 4.1: The average and best classification accuracy of the proposed *BBLBP* operator based universal multi-view facial expression recognition system.

Table 4.1 shows that the average performance of the proposed facial expression recognition system based on *BBLBP* feature is 67.66%, the best performance is achieved using the block based local binary pattern representation extracted at the scale of 3. However, when compared with the *BBLTP* based system described in Chapter 3, this system underperforms by about 9%.

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	64.23	10.16	5.01	2.79	14.70	3.11
Disgust	9.85	69.52	6.49	3.90	6.31	3.93
Fear	5.35	7.42	63.46	9.71	8.14	5.91
Happy	2.21	3.66	7.54	83.84	1.59	1.16
Sad	14.87	5.40	7.29	2.74	67.09	2.60
Surprise	2.19	3.63	4.36	1.70	2.78	85.34

Figure 4.5: The classification confusion matrix of the *BBLBP* feature based system for classification of 6 prototypic facial expressions.

By studying Figure 4.5, it can be concluded that *fear* and *anger* are the most difficult facial expressions to recognize for the proposed system, while *anger* is most confused with the *sadness* facial expression, with a classification accuracy of 64.23% and *fear* is misclassified with all other expressions with classification accuracy of 63.46%. The facial expressions of *surprise* and *happiness* are the easiest to recognize by the system with a classification accuracy of 85.34% and 83.84% respectively. *Disgust* and *sadness* are classified slightly better than *anger* and *disgust*, but the accuracy is much lower than those of facial expression of *surprise* and *anger*.
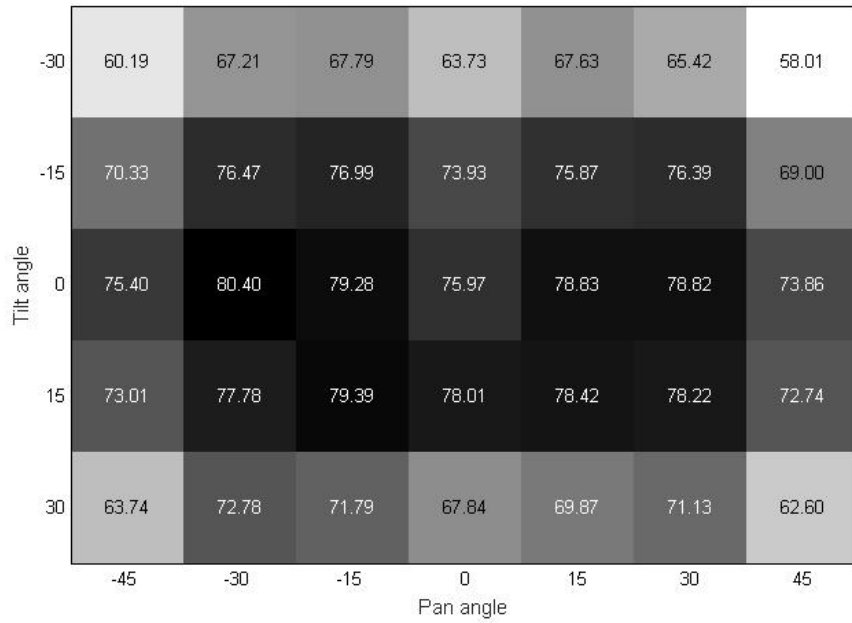
(a)
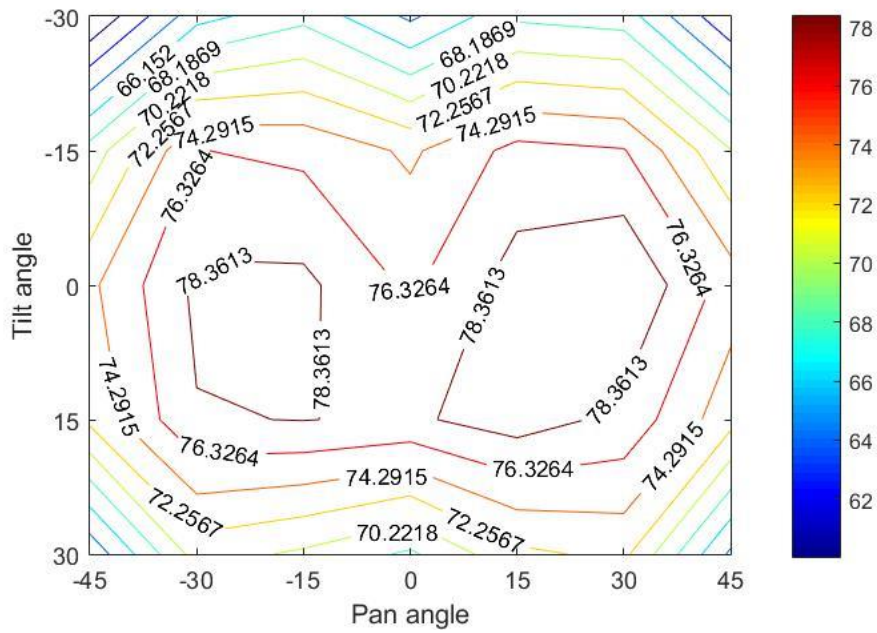
(b)

Figure 4.6: (a) The matrix of classification accuracy for facial expression images with various combinations of tilts and pan angles; (b) a contouring map based on the classification accuracy.

Figure 4.6 reveals that the performance of the proposed facial expression recognition system is significantly affected by the changes of head pose, and especially that the tilt angle exerts more influence than the pan angle. Considering the influence of the tilt angle, a negative tilt angle affects the overall classification accuracy more than a positive tilt angle. And the system performance is almost symmetrically reflected, with a sum of difference of 15% on positive and negative pan angles, and the negative pan angles performing better than the positive pan angles. Furthermore, it is observed in Figure 4.6 that the pan angle of $0°$ is not the optimal for facial expression classification because a higher classification accuracy is obtained at pan angles of $\pm 15°$ and $\pm 30°$.
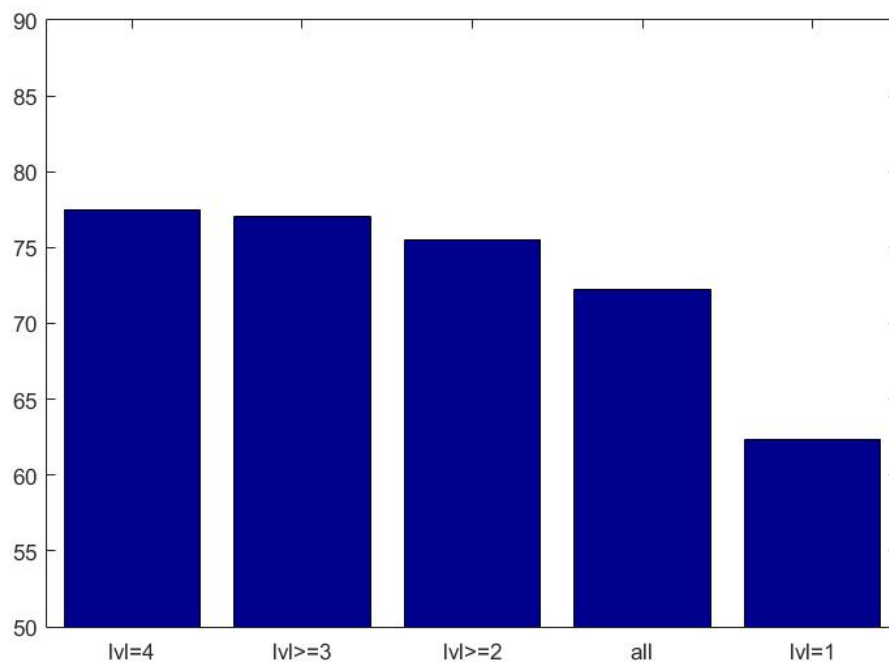
Figure 4.7: The classification accuracy of the *BBLBP* based system with respect to the change in intensity level (*lvl*) of the facial expression.

As shown in Figure 4.7, the classification accuracy of the proposed system does not significantly change before facial expression images of intensity level of 1 are added in to the classification. The classification accuracy for the proposed system operating

at intensity level of 1 is 62.37%, which is about 15% lower than the accuracy obtained at the highest intensity level.

Although the *BBLBP* based system has achieved outstanding performance, when compared with *BBLTP* based system, it underperforms by about 9%. The *BBLBP* representation is in general less effective in describing all facial expressions for universal multi-view facial expression recognition task, and the *BBLTP* representation is significantly better in reducing confusions of *anger*, *disgust*, *fear,* and *sad*.

4.4.2 Multi-scale uniform local binary pattern with block based feature extraction ($BBLBP^{ms}$)

To obtain a feature representation using the block based multi-scale local binary patterns ($BBLBP^{ms}$), the LBP^{ms} operators with 8 different scales of extraction are applied to a facial expression image. The local binary patterns extracted from all scales are filtered using the F-score feature selection algorithm, and by ranking the F-scores the same number of the most distinctive features are selected for each scale.

As shown in Table 4.2, the proposed system produced an overall classification accuracy of 73.52%. This system slightly outperforms the best configuration of *BBLBP* feature based system by 1.27%, and it is about 9% worse than the $BBLTP^{ms}$ based systems as described in Chapter 3.

	Accuracy
$BBLBP^{ms}$	73.52%

Table 4.2: The performance of the established universal multi-view facial expression recognition system using $BBLBP^{ms}$ with F-score feature selection.

Figure 4.8: The classification confusion matrix obtained for the proposed $BBLBP^{ms}$ feature based system.

Figure 4.8 demonstrates that the facial expressions surprise and *happiness* are best classified by the proposed system, while *fear* is the most confused facial expression. The *fear* expression is misclassified with all other expression to the same extent, and *anger* and *sadness* are misclassified the most among all pairs of facial expressions. *Disgust* is misclassified with *anger* and *fear* by the proposed system. Combining confusion matrices of both local binary pattern and the multi-scale local binary patterns in Figures 4.5 and 4.8, it can be concluded that these two operators are highly related, with similar misclassification trends and similar performance.
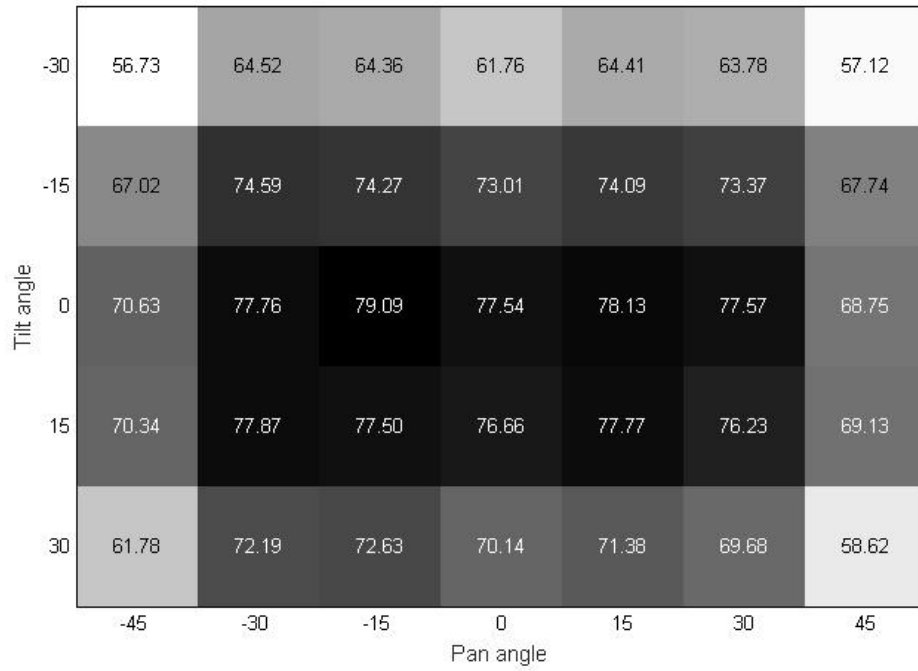
(a)
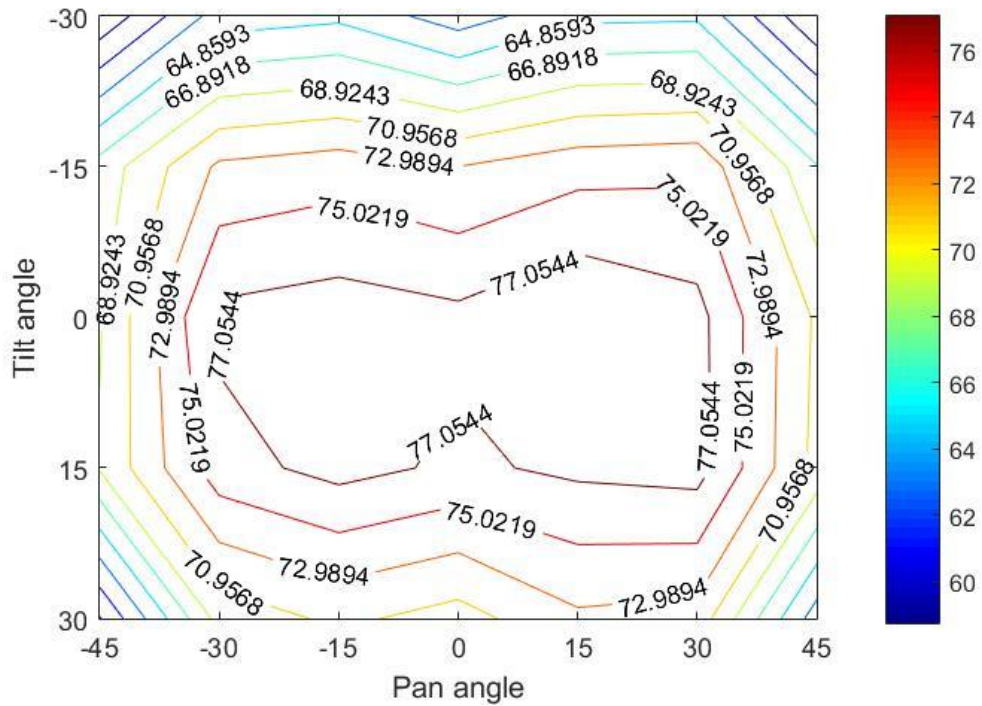
(b)

Figure 4.9: The change of the classification accuracy of the proposed system using $BBLBP^{ms}$ w.r.t. the tilt and pan angles.

Figure 4.9 shows that that the pan angle of $\pm 15°$ is the optimal for facial expression recognition, with an average performance of 76.36%. In addition, tilt angle has the most significant influence on the system performance when compared with the pan angle. According to the performance obtained in this experiment, it is shown that facial expression recognition is nearly symmetrical on both sides of the face. The best viewpoint is found at a tilt angle of 0° and pan angle of 15°.
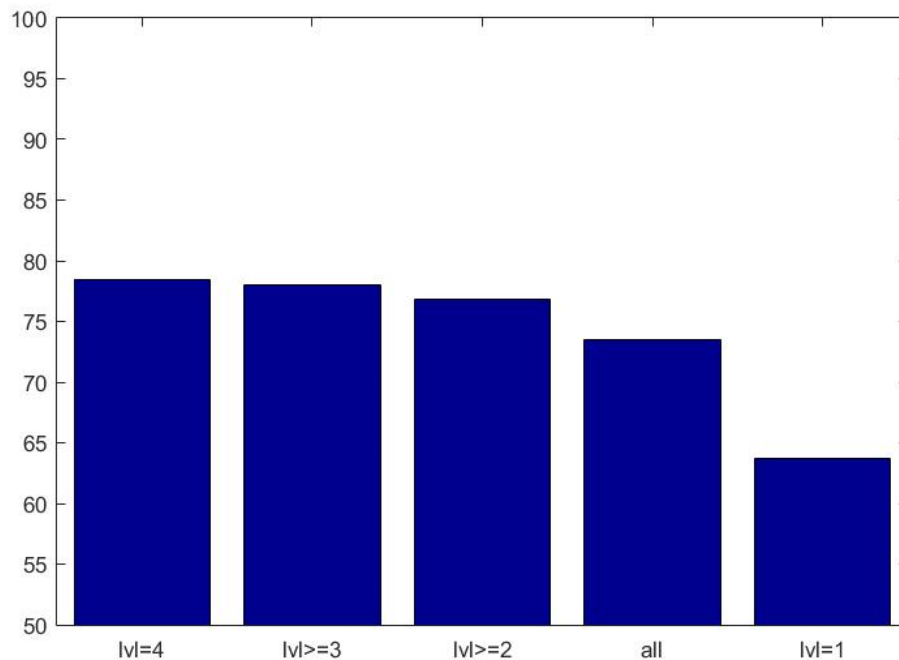
Figure 4.10: The system performance in classification of six universal facial expressions at various intensity levels using $BBLBP^{ms}$.

As seen in Figure 4.10, the multi-scale local binary pattern based approach degrades in performance as lower intensity levels are included, and shows the worst performance at the lowest (1) intensity level. The recognition difference between the

proposed system operating on highest and lowest intensity is 14.7%, where 63.37% is observed at the lowest intensity level and 78.47% at the highest intensity level.

Comparing experimental results between $BBLBP^{ms}$ based system and $BBLTP^{ms}$ based system described in Chapter 3, it can be observed that $BBLBP^{ms}$ based system performs generally worse than $BBLTP^{ms}$ based system. The $BBLBP^{ms}$ based representation increase the confusion of *anger*, *disgust*, *fear*, and *sad*. As a result, the overall performance of the $BBLBP^{ms}$ based system is affected.

4.4.3 Histogram of oriented gradients (HOG)

In this experiment, the histogram of oriented gradients (HOG) operator is applied densely with an overlap of half blocks size to extract the HOG representation for the facial expression image. All densely extracted HOG features are concatenated into one feature vector as the final representation for the query image. To reduce the dimensionality of the HOG representation, the F-score based feature filtering is implemented to select 2,500 features from the complete HOG features with the best F-score. Finally, a 10-fold cross-validation is conducted. The average classification accuracy of the established system is 67.33% as presented in Table 4.3.

	Accuracy
HOG	67.33%

Table 4.3: The classification accuracy of the established universal multi-view facial expression recognition system using the HOG feature is shown.

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	62.24	7.90	5.44	3.17	17.47	3.78
Disgust	11.42	59.62	8.47	6.58	6.81	7.09
Fear	7.16	9.66	52.82	13.51	8.87	7.97
Happy	3.29	4.85	8.42	80.16	1.62	1.66
Sad	17.46	5.22	5.91	1.94	65.83	3.64
Surprise	2.55	3.92	3.99	2.49	3.74	83.31

Figure 4.11: The confusion matrix of the proposed facial expression recognition system using the *HOG* feature representation.

As shown in Figure 4.11, for the *HOG* features-based facial expression recognition system, *fear* is the most misclassified expression at 52.82% and *surprise* and *happiness* the best at 83.31% and 80.16% respectively. The facial expressions *sadness* and *anger* are misclassified with each other, as are *disgust* and *anger*. *Happiness* is misclassified with *fear* the most. *Disgust* ranks third in classification accuracy.
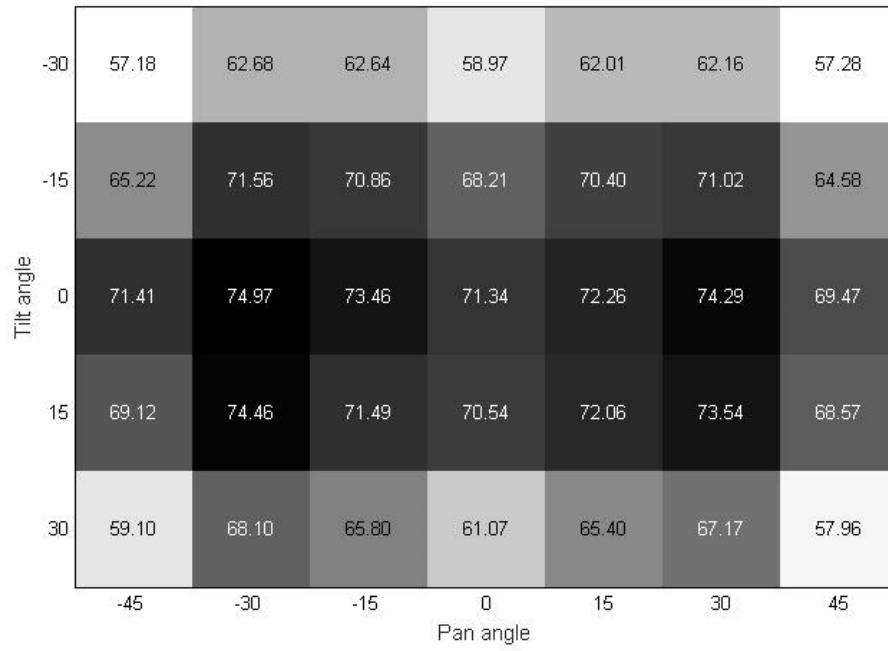
(a)
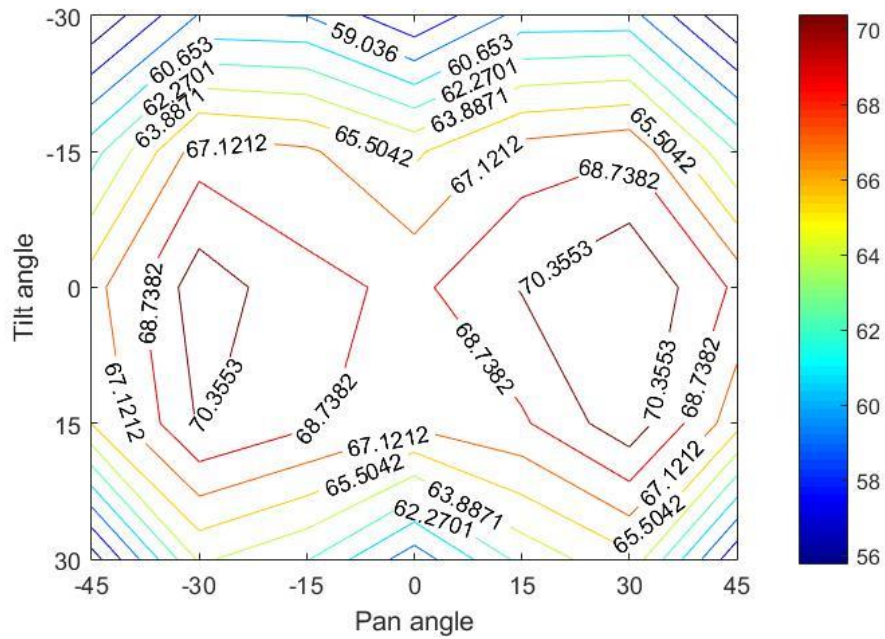
(b)

Figure 4.12: (a) The classification accuracy matrix for various views of different combinations of tilt and pan angles; (b) contour map of classification accuracy of the proposed systems at various views.

Among all combinations of head poses, the best classification accuracy is obtained at the tilt angle of 0° and pan angle of−30° according to the experimental results shown in Figure 4.12 (a). The Figure strongly indicates that an increase of tilt angle either positively or negatively will reduce the overall performance of the system significantly, and a negative change of tilt angle has more significant influence on the system performance compared to a positive increase. A detailed trend of the pan and tile angle's influence over the performance of the proposed system is reflected in the contour map in Figure 4.12 (b). In addition, an angle of ±30° is observed as the best pan angle for facial expression recognition using the histogram of oriented gradient feature representation.
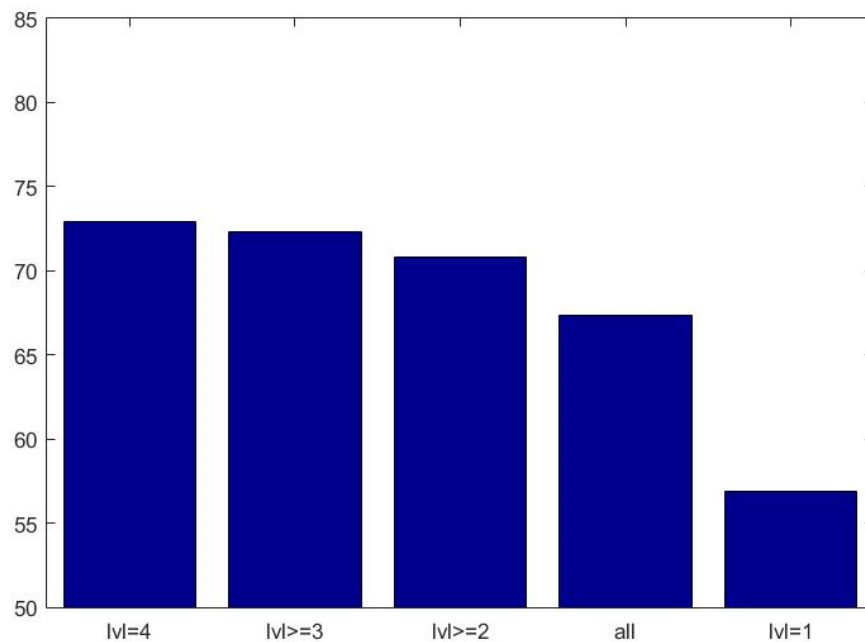
Figure 4.13: System performance difference using the histogram of gradient feature.

As shown in Figure 4.13, the classification accuracy of the universal multi-view facial expression recognition system using the HOG feature is 72.89% on the highest intensity level, and 56.88% on the lowest intensity level. The overall classification accuracy is reduced by the poor classification accuracy the system has managed to achieve at the lowest intensity.

4.4.4 Gray level co-occurrence matrix ($GLCM$)

In this experiment, the block-based feature extraction of gray level co-occurrence matrix ($GLCM$) is carried out. First of all, an intensity level reduction is performed which reduces the total number of intensity levels of the image to 8 levels. Secondly, the input image is divided into 6×4 blocks. Then, the operator is applied to calculate the gray level co-occurrence matrix. After that, the calculated matrix is transformed into vectors, and eventually concatenated into one vector as the feature representation.

To thoroughly investigate the potential of the gray level co-occurrence matrix as a representation for universal multi-view facial expression recognition, 4 spatial relations are examined, which produce 4 gray level co-occurrence matrices, at $0°$, $45°$, $90°$, and $135°$, with a distance of 1.
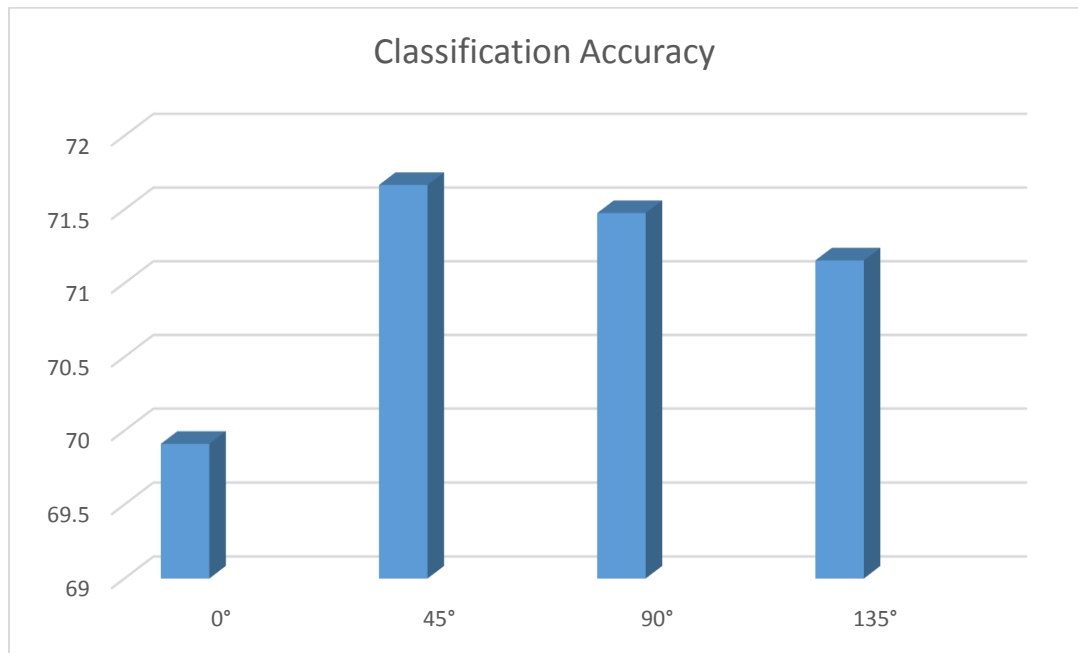
Figure 4.14: The classification accuracy for gray level co-occurrence matrices calculated at $0°$, $45°$, $90°$, and $135°$, with a distance of 1.

It is observed from Figure 4.14 that $GLCM$ calculated at $0°$ has the worse performance at 69.92% while the $GLCM$ for $45°$ yields the best performance at 71.48%, which suggests that a spatial relation of $45°$ with immediate distance of 1 has the highest distinctiveness as a representation for universal multi-view facial expression recognition.

In addition, the $GLCMs$ of 4 spatial relations are fused together by adopting the F-score feature selection described in Section 3.5. For each $GLCM$, the dimensionality of the feature is reduced to 625. As a result, the total dimension of the final feature representation after fusion is $625 \times 4 = 2500$. With the 10-fold cross-validation, the $GLCMs$ achieves an overall 72.08% classification accuracy as shown in Table 4.4, and the confusion matrix is as shown in Figure 4.15.

	Accuracy
GLCMs	72.08%

Table 4.4The classification accuracy with the combined *GCLMs*.

Figure 4.15: The confusion matrix of the proposed universal multi-view facial expression recognition system using four spatial *GLCMs*.

By studying the confusion matrix shown in Figure 4.15, it can be concluded that *disgust* and *fear* are the most difficult expressions to classify for the proposed system. *Anger* is misclassified with *sadness* the most and *disgust* the second most while *fear* is misclassified with *disgust* the most, and *happiness* the second most. Also, a classification accuracy difference of 34.51% is observed in this experiment, demonstrating that the classification accuracy among various facial expressions varies greatly. *Surprise* and *happiness* are classified with the highest accuracy by the proposed system.
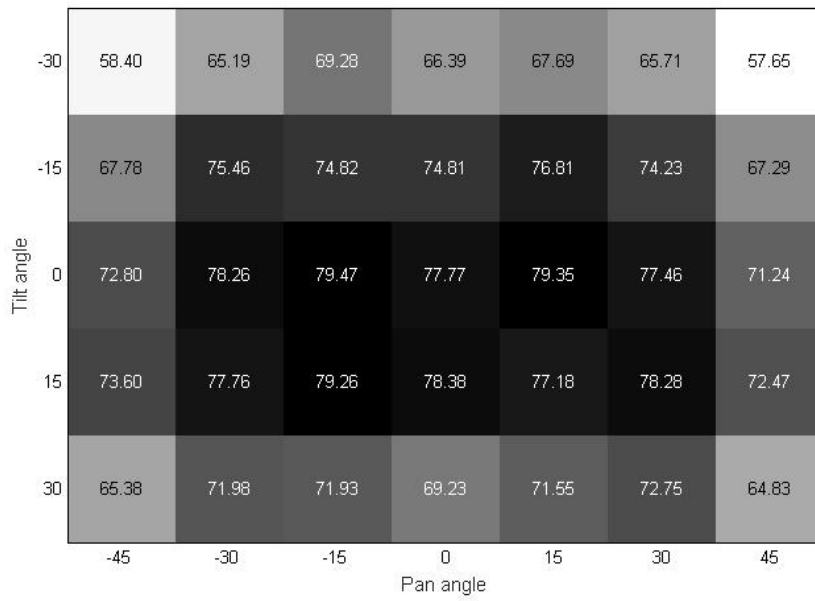
(a)
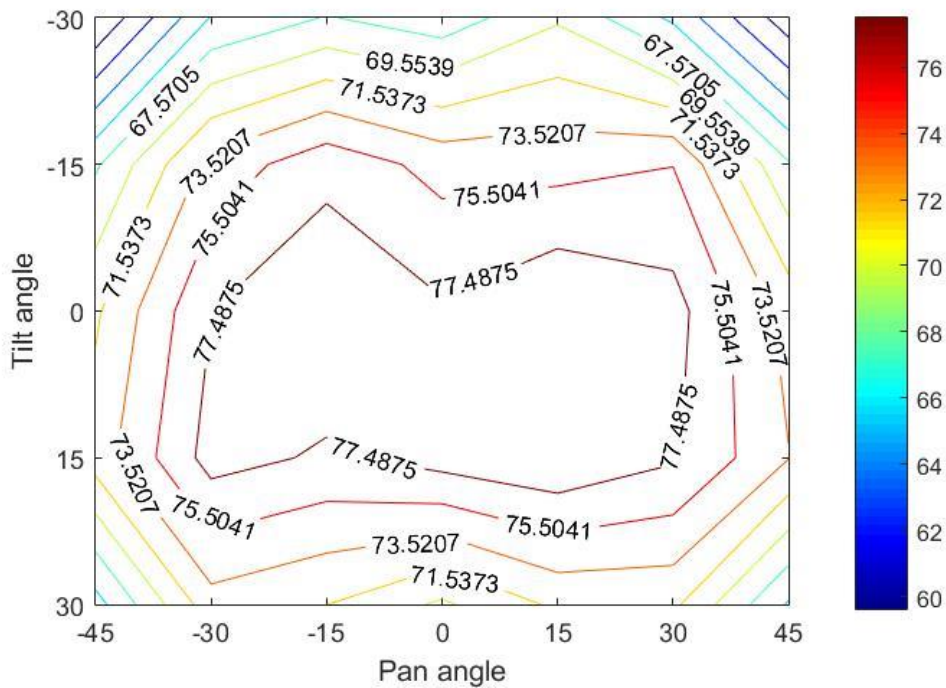
(b)

Figure 4.16: The classification accuracy for the proposed universal multi-view facial expression recognition system using four spatial $GLCMs$: (a) classification accuracy

yields at each view; (b) contour map of the classification accuracy with respect to variation of pan and tilt angle.

It is observed in Figure 4.16 that the optimal pan view for facial expression recognition for the proposed facial expression recognition system is $\pm 15°$, and the best system performance is observed at pan angle of $-15°$ and tilt angle of $0°$. Also, it is seen that facial expression recognition from the left side view slightly outperforms recognition from the right side. Furthermore, the results shown in the Figure strongly indicate that tilt angle has a significant influence over the performance of the system, and, more specifically, that a negative view angle affects the performance the most.
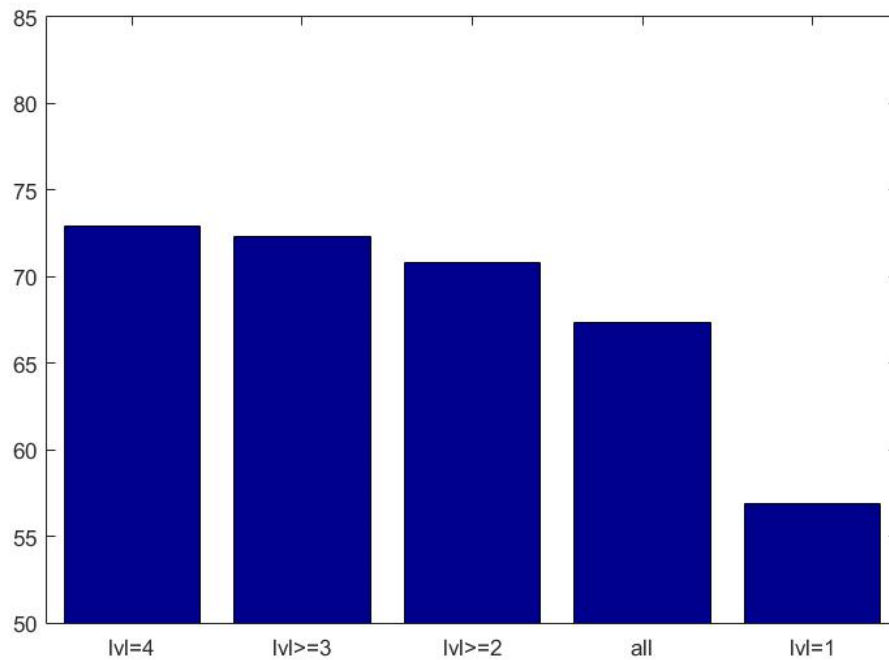
Figure 4.17: The performance of the proposed system in classification of six universal facial expressions at various intensity levels with $GLCMs$.

Figure 4.17 shows that the classification accuracy of this system diminishes at the same rate from intensity 4 to 2, and then falls significantly to 61.78% at an intensity level of 1. The overall classification accuracy is obviously reduced by the poor performance the system yielded at the lowest intensity level. The classification difference observed between the highest and lowest intensity levels is approximately 16%.

4.4.5 Level of difference descriptor (LOD)

To extract the level of difference descriptor, the block-based 2-layer feature extraction with an overlap of a half block is applied as illustrated in Figure 4.18. The size of the basic extraction block that the level of difference descriptor is operating on is 21×24 pixels. The first layer consists of 6×4 basic extraction blocks while the second layer has 5×3 blocks. The LOD features extracted from all blocks are concatenated to form the final representation. The length of the resulting LOD feature is 624.
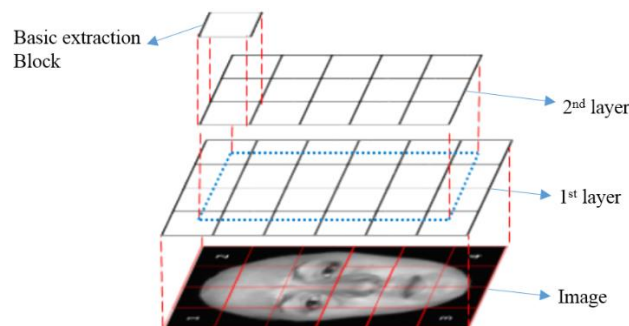
Figure 4.18: The feature extraction process for the LOD descriptor is illustrated.

	Accuracy
LOD	48.92%

Table 4.5: The overall performance of the LOD descriptor for the proposed facial expression recognition system

Employing the feature representation generated by the *LOD* descriptor, the proposed universal multi-view facial expression recognition system classifies 6 prototypical facial expressions with an accuracy of 48.92%, as shown in Table 4.5.

By studying the confusion matrix, as illustrated in Figure 4.19, it is seen that the *LOD* descriptor cannot code the distinctiveness of the facial expressions *disgust* and *fear* which are seen to be poorly classified, and the classification accuracy is similar to a random guess. However, the proposed system can classify the expression *surprise* at a classification accuracy of 84.64% and *happiness* at 75.52%. *Sadness* and *anger* are the most confused classes except for the expressions of *disgust* and *fear*.

	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	55.29	4.90	3.31	13.51	15.30	7.69
Disgust	17.26	21.17	6.57	23.49	15.31	16.19
Fear	12.81	8.14	14.54	29.52	17.59	17.39
Happy	6.26	4.49	3.79	75.52	5.69	4.24
Sad	27.94	4.84	4.45	10.74	42.36	9.69
Surprise	3.09	2.61	2.03	3.68	3.96	84.64

Figure 4.19: The classification confusion matrix for the proposed system.

As shown in Figure 4.20, the optimal pan view for facial expression recognition is observed at $\pm 15°$. Comparing the difference of influence between pan and tilt angles, the tilt angle has more significant impact on the performance of the proposed facial

expression recognition system. In addition, a negative tilt angle affects the classification accuracy substantially. And, the best recognition view for this system is observed at a tilt angle of 15° and pan angle of 15°.
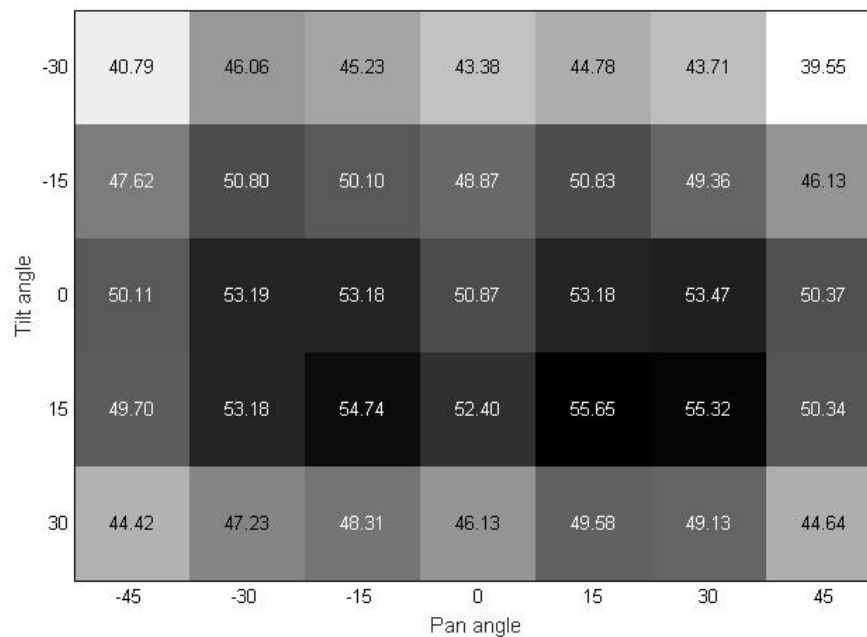
Figure 4.20: The classification accuracy matrix from various camera views.

4.4.6 LOD descriptor as a supplement to other texture descriptors

As noted earlier, the level of difference descriptor is suggested as a supplement for those local texture descriptors, which are weak or are not capable of coding the appearance of the facial expression image. The experiment presented in this section thoroughly investigates its influence as a supplement in terms of classification accuracy when the *LOD* descriptor is employed together with the block-based multi-

scale local binary pattern ($BBLBP^{ms}$) operator, block-based multi-scale local ternary pattern ($BBLTP^{ms}$) operator, four directional gray level co-occurrence matrices ($GLCMs$), and the histogram of oriented gradients (HOG) operator.
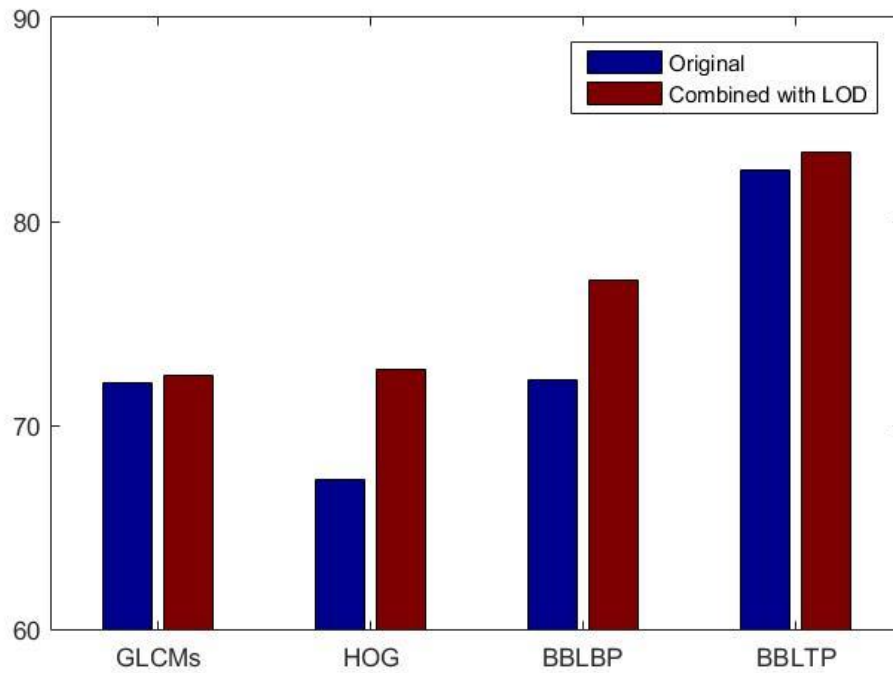
Figure 4.21: Classification accuracy of $GLCMs$, HOG, $BBLBP^{ms}$, and $BBLTP^{ms}$, and their performance when used with the LOD descriptor.

As shown in Figure 4.21, combining the LOD feature with features generated by other texture descriptors can generally improve the classification accuracy when the same numbers of features are used. Also, when employing LOD with the HOG feature and the $BBLBP^{ms}$ feature, the classification accuracy of the proposed system is improved substantially, specifically 4.9% for $BBLBP^{ms}$ and 5.42% for HOG.
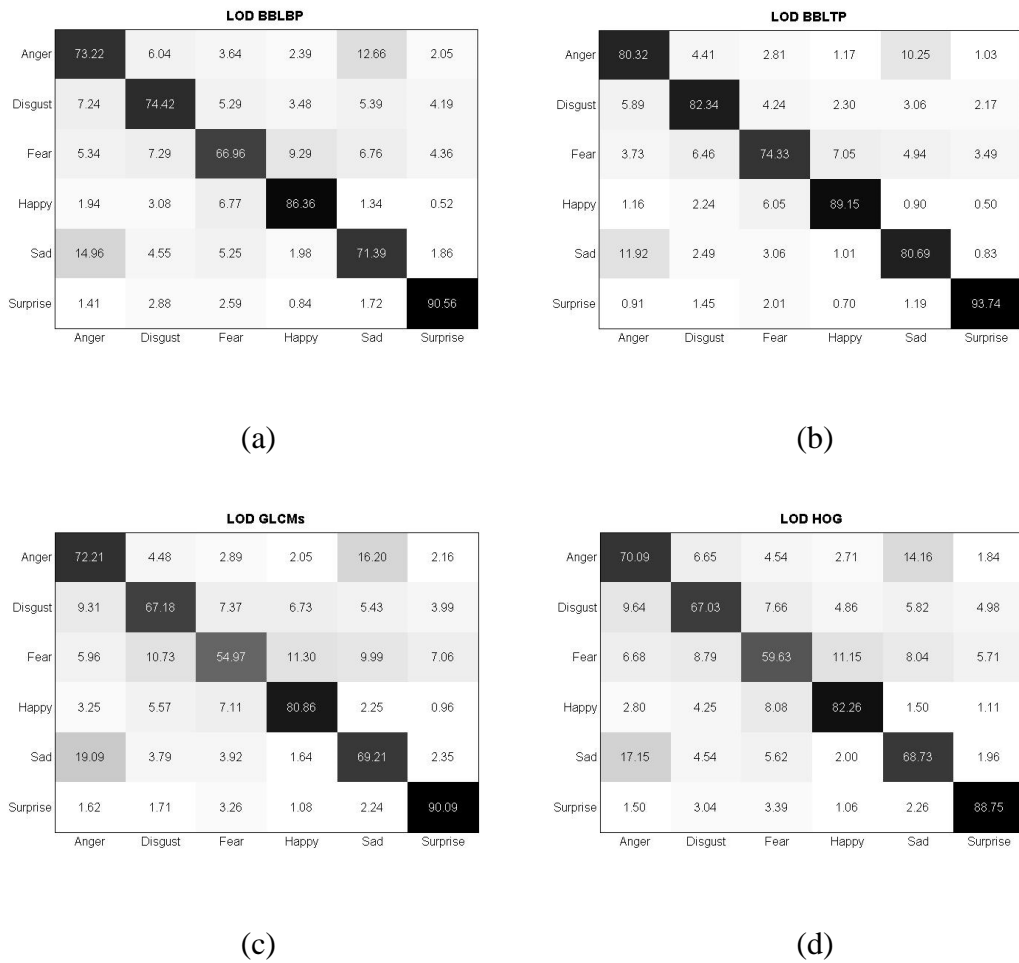
Figure 4.22: The classification confusion matrix for the combined representation of LOD with (a) $BBLBP^{ms}$, (b) $BBLTP^{ms}$, (c) $GLCMs$, and (d) HOG.

Considering the combination with $BBLBP^{ms}$ feature, it is observed in the confusion matrix shown Figure 4.22 that the classification accuracy for all six expressions is improved in general, with the largest improvement observed for the *anger* expression of around 10%, while for other expressions an average improvement of around 5% is delivered. And it is seen that the *LOD* features help to reduce the confusion between the facial expressions *anger* and *sadness*. The least improvement is observed for *fear* and *happiness*.

Regarding the combination with $BBLTP^{ms}$ feature, the classification accuracy for each pair of facial expressions remains unchanged, except the recognition accuracy for *disgust* and *sadness* is improved by about 3%, and the classification accuracy for *fear* drops by about 1%.

Comparing the confusion matrix relating to HOG features, it is observed that the overall confusion between all pairs of facial expression is improved. While the least improved is observed for *happiness* (by about 2%), the classification accuracy of the proposed system has improved by around 7% on average

By studying the changes in the confusion matrix relating to $GLCMs$, it is observed that the general confusion between pairs of facial expression is not changed substantially.

(a)   (b)

(c)   (d)
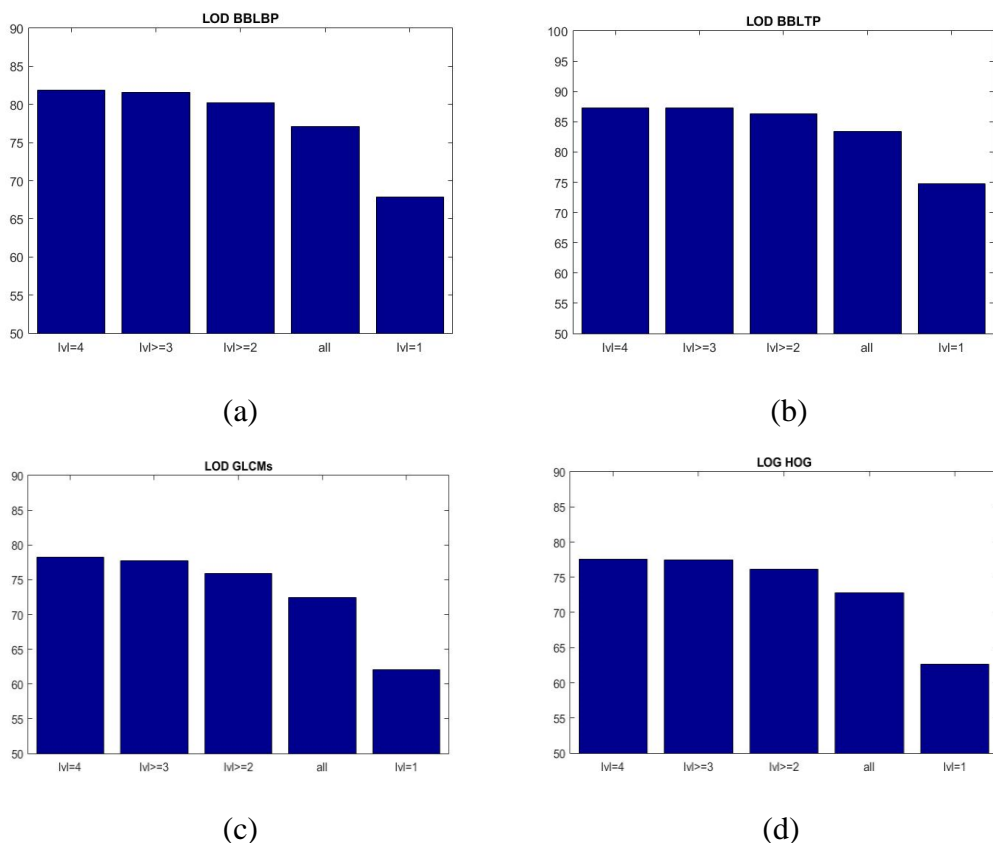
Figure 4.23: The classification difference with respect to intensity level of the proposed classification system when the LOD feature is combined with other features: (a) $BBLBP^{ms}$, (b) $BBLTP^{ms}$, (c) $GLCMs$, and (d) HOG.

As observed in Figure 4.23, the feature representations that combine the LOD feature show a similar trend, namely that the classification accuracy of the proposed system does not degrade significantly until the facial images of the lowest intensity are included in the scope of the classification. Additionally, combining the LOD feature with HOG and $BBLBP^{ms}$ not only stabilizes the classification accuracy at the highest two intensity levels, but also improves the classification accuracy of the new combined feature (one is constructed of LOD and HOG feature, and the other LOD and $BBLBP^{ms}$ feature) based systems in general. The average performance difference between the highest and lowest intensity levels is about 15% for all combined features except when the $BBLTP^{ms}$ feature was included, the difference is around 13%. The lowest difference occurs when the LOD feature is included, which implies that a fused feature that includes the LOD feature can reduce the performance difference observed between the highest and lowest intensity levels. In addition, it is revealed in Figure 4.23 that facial expression recognition using all these combined features at the lowest intensity is significantly poorer than for classification carried out at higher intensity levels, which then decreases the overall performance of the system and restricts its usability.

4.4.7 Fusion of state-of-the-art texture features

To further investigate a search for the most suitable representation for the proposed universal multi-view facial expression recognition system, the full combination of the aforementioned texture descriptors (except the level of difference descriptor) is examined on a pairwise basis to form a representation for the proposed recognition problem. The texture features that are selected for this experiment can be extracted on a real time basis which means a single extraction takes less than 33ms. Moreover, to manage the memory requirement for this experiment, the total number of features employed is restricted to 2,500, the number with which the maximum dimensionality

that our system hardware (i.e. our system has a RAM of 16 GB) can stably operate, for the combined feature representations. For different pairs of features, a specific ratio is calculated which determines the number of features to select from each type of feature. The ratio is calculated using Equation 4.2.

$$n = \frac{2500}{P} w \qquad\qquad 4.2$$

where n is the total number of features that need to be selected; P is the summation of classification accuracy for both original texture descriptors; w is the original classification accuracy of the features to be selected.
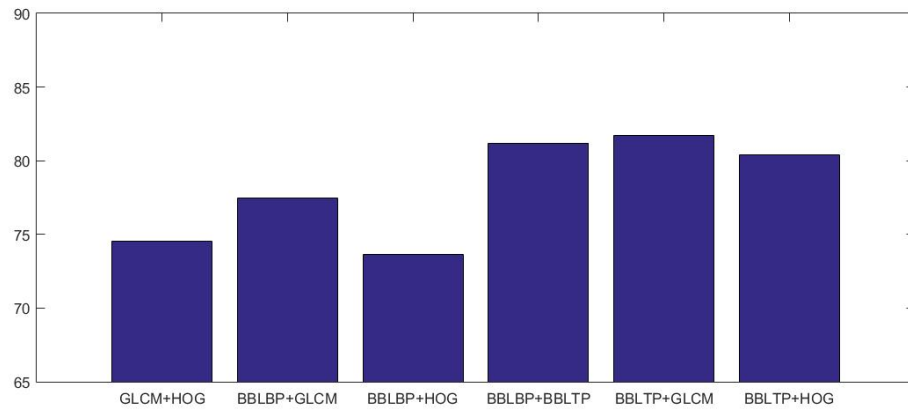
Figure 4.24: The classification accuracy for each pair of texture descriptors based on: $GLCMs, HOG, BBLBP^{ms}$, and $BBLTP^{ms}$.

As shown in Figure 4.24, the best classification accuracy for the proposed system is achieved with the combined feature representation of $BBLTP^{ms}$ and $GLCMs$ at 81.725%, followed by the combined feature representation block- based local binary pattern and local ternary pattern at 81.2%. And it is seen that all the top three system performances recorded are achieved when the local ternary pattern is fused in the feature representation. Most combined representations have outperformed a feature representation generated with a single texture descriptor, except local ternary patterns. It is revealed in the experimental results that the local ternary pattern is more efficient and informative in the description of the difference of six prototypic facial expressions. When combined with other texture features, the overall performance is lower than using $BBLTP^{ms}$ feature alone. It is also observed that combining the histogram of oriented gradient feature with other features provides no improvement at all in terms of classification accuracy for the proposed system setup. In addition, the overall system performance slightly increases (by around 3%) when the combined $BBLBP^{ms}$ and $GLCMs$ feature is employed.

It is strongly indicated in Figure 4.25 that that *surprise* and *happiness* are classified most accurately by the proposed universal facial expression recognition system while *fear* is the most misclassified expression. *Anger* and *sadness* are misclassified with each other substantially.

(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.25: The classification confusion matrix for the proposed facial expression recognition system using feature representations of combinations of all pairs of texture features: (a) $BBLBP + BBLTP$; (b) $BBLBP + GLCMs$; (c) $BBLBP + HOG$; (d) $BBLTP+GLCMs$; (e) $BBLTP+HOG$; (f) $GLCMs+HOG$.
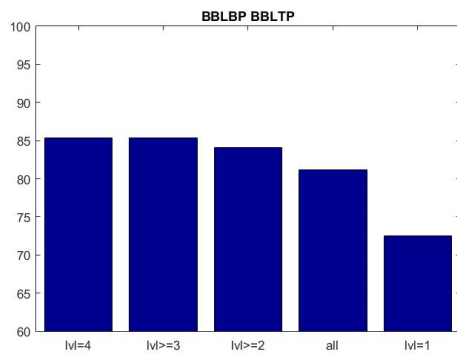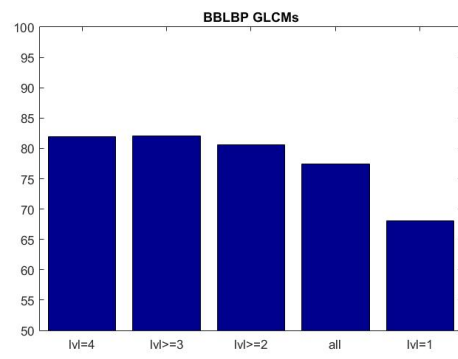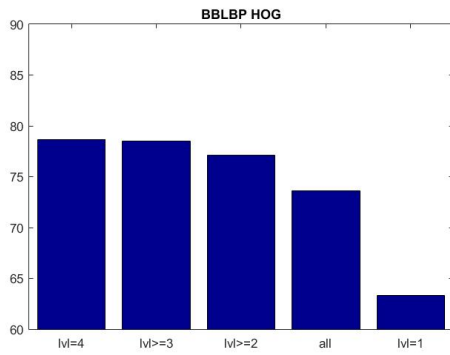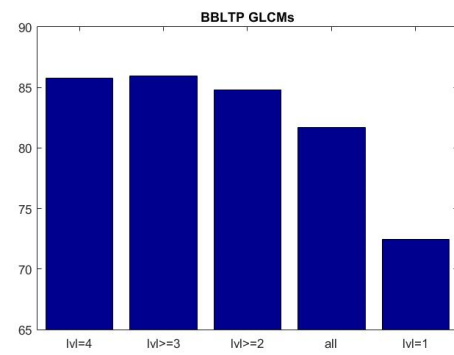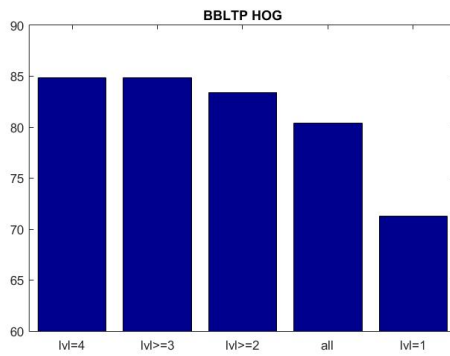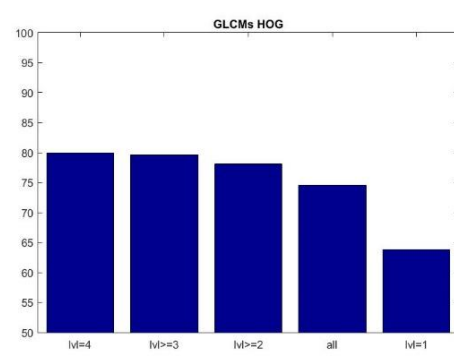
Figure 4.26: The classification of six prototypic facial expressions with increased number of intensity levels (*lvl*) using combined features: (a) $BBLBP+BBLTP$; (b) $BBLBP+GLCMs$; (c) $BBLBP+HOG$; (d) $BBLTP+GLCMs$; (e) $BBLTP+HOG$; (f) $GLCMs+HOG$.

It is demonstrated in Figure 4.26 that the classification accuracy for all proposed system setups declines as the intensity of the facial expression decreases, which confirms that facial expressions with lower intensity levels are difficult to classify in general. It is seen that classification of expressions at the lowest intensity level is the most difficult for the proposed system. The difference of classification accuracy in terms of accuracy between the highest and lowest intensities is roughly grouped into two parts. The combined features of $BBLBP^{ms}$ and $GLCMs$, $BBLTP^{ms}$ and HOG, and $BBLBP^{ms}$ and $BBLTP^{ms}$, show a difference of approximately 13% while this is around 15% for the others. In addition, according to the experimental results, it can also be concluded that the performance of the proposed system can be divided into two parts: one without the lowest intensity level and the other including the lowest intensity. For all combined features, the performance of the system across the highest three intensity levels varies less than 2%, and is significantly better than the recognition accuracy observed at the lowest intensity. The recognition accuracy observed at the lowest intensity can be considered as a lower bound of performance that a facial expression recognition system can achieve.

To sum up, the findings obtained by studying and analysing the experimental results can be summarized as follows:

First, considering all the experimental results obtained using the *BU-3DFE* database that is obtained in this series of experiments, it is strongly suggested that the block-based local ternary pattern is the most efficient and effective texture feature for the universal multi-view facial expression recognition among all other state-of-the-art texture features that have been investigated in this study, and its performance is stable and outperforms the combined feature representations constructed from state-of-the-art texture descriptors, for universal multi-view facial expression recognition system in terms of classification accuracy at 83.23% for 35 head poses across all different intensity levels included in the database.

Second, reviewing the results from the head pose perspective, it is observed that tilt angle is the most significant factor that affects the classification accuracy. Especially, a camera view that captures the input images from a position above the subject dramatically increases the overall recognition difficulty for the proposed system. Also, it is reflected in all the experimental results that the pan view of 0° is not the best pan view for facial expression recognition, and the best recognition accuracy is often observed at pan angle of $\pm15°$ or $\pm30°$ depending on the feature representation and system configuration adopted. In addition, it is shown in the experiment results that there is not a significant difference in universal recognition of facial expression between a situation where performed from the left side and right side, which strongly suggests that the facial expression is symmetrical with respect to the proposed system in the universal classification task.

Third, it is found that the level of difference descriptor can be used as a supplement when combined either with the histogram of oriented gradient feature and the block-based local binary pattern feature, where the classification accuracies for the combined feature is significantly improved to 72.75% and 77.15% for combination with HOG and $BBLBP^{ms}$ respectively when the same dimensionality of the feature representation is employed. Especially, it should be noted that the performance which the combined feature of LOD and $BBLBP^{ms}$ has achieved is significantly better than that reported in a similar study by Moore and Bowden [166] in which the combined feature of multi-scale local binary pattern and local Gabor binary pattern is used to devise a multi-view facial expression recognition system using the same database. The total number of feature dimensions that are utilized in our study is about $\frac{1}{72}$ of the feature dimensionality of Moore and Bowden's system, and the performance that the LOD and $BBLBP^{ms}$ based system has achieved in this study outperforms $BBLBP^{ms}$ and local Gabor binary pattern based system by about 6% on a more difficult setup in terms of numbers of views and facial expression intensities.

Fourth, according to the experimental results obtained in the feature representation fusion study, it is observed that not all combinations of features can be helpful in improving the overall classification accuracy with a restricted length of feature representation employing the F-score based feature selection strategy. An inappropriate fusion can degrade the system performance as well. For example, the HOG feature when used with $BBLBP^{ms}$ and $BBLTP^{ms}$ features reduces the overall classification accuracy of the system.

Finally, it is revealed by the reported experiment results, in an exploration of the intensity level's influence over the classification accuracy, that the universal multi-view facial expression recognition at the lowest level is the most challenging for all proposed systems and degrades the overall performance and usability of a facial expression recognition system dramatically. The performance difference between a system with or without inclusion of the intensity level of 1 is large, which strongly suggests that the facial expression classification at the lowest level (i.e. the onset phase of an expression) can be used as an evaluation criterion for examining the performance, robustness, and usability of a facial expression recognition system.

4.5 Conclusion

In this chapter, an exploration of the feature representation for universal multi-view facial expression recognition using state-of-the-art local descriptors is conducted, and the corresponding experimental results are thoroughly analysed and discussed in relation to practical applications. Importantly, a novel level of difference descriptor is also introduced as a supplement to use with other state-of-the-art texture features, where a promising improvement of overall classification accuracy is observed.

In conclusion, with this comprehensive study, it is strongly suggested that adopting the proposed combined feature, which is generated from the fusion of features extracted by various local descriptors, might not always achieve a better classification accuracy. Besides, the proposed novel LOD descriptor can be used as a supplement

for local binary pattern and histogram of oriented gradient features to improve the overall classification accuracy of a facial expression recognition system.

In the following Chapter 5, a novel categorization strategy for facial expressions is proposed in order to extend the application of facial expression recognition into more practical day-to-day scenarios.

Chapter 5

Subcategories of facial expressions for practical applications

This chapter reviews the fundamental theories of emotions and further explains the key dimensions in the emotional space, and then proposes a set of novel categorization methods for facial expressions to be used in the design of an automatic facial expression recognition system. In addition, a series of experiments is reported which inspect the influence, which the novel categorization brings to our universal multi-view facial expression recognition system. A detailed presentation of experimental results and analysis is also included at the end of this chapter.

Section 5.1 explains the general background and motivation of this study and Section 5.2 reviews the fundamental theories of emotions and affective spaces. In Section 5.3, the methodology of developing a novel grouping of facial expressions is presented. And Section 5.4 describes thoroughly a series of experiments that is conducted in this study from experimental setups of our preliminary research to detailed analysis of the experimental results. Finally, in Section 5.5, a brief summary of this chapter's content is presented.

5.1 Introduction

As a "hot topic" within the computer vision and machine learning community, facial expression analysis has been intensively studied in recent years. However, most of the research and development of automatic facial expression recognition systems has been conducted based on a fundamental acknowledgement that emotions are discrete, and, thus, facial expressions reflect this. In the light of this fundamental understanding of emotion and facial expression, many researchers in the field of computer vision and pattern recognition [90], [25], [167], [168], [169] have attempted to classify facial expressions automatically into one of the prototypic facial expression families, consisting of expressions defining the emotions of *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* [12]. Other researchers [170], [171], [55], [102], [103], [28] have employed the facial action coding system (*FACS*), which is generalized by Ekman and Friesen [6], and devised an automated recognition method to analyse the facial actions in various region of the face in an expressive facial image, and then classify the facial expressions based on the protocol suggested by the facial action coding system.

Despite the significant progress in the automatization of facial expression analysis, the practical applications of automatic facial expression recognition have advanced considerably less than they otherwise might have due to some practical difficulties in the application of such a technology. First of all, the spontaneous facial expressions typically occurring in daily life are generally less pronounced than the "posed" expressions that are usually captured in the laboratory [114]. This, as a result, brings about substantial difficulty in recognizing facial expressions in practical application scenarios (see also our previous findings reported in Chapter 3 and 4, that recognizing facial expressions at low intensity level is challenging in general and restricts the general usability and robustness of a facial expression recognition system). Secondly, there is a deficiency in the standardized facial expression data for devising practical and universal facial expression recognition system. Data availability is one of the most influential factors that affects a recognition solution because recognitions system is built based on these exploratory data collected at first hand. Facial expression data

collected in laboratory conditions or synthetic facial expression data theoretically would lead to a suboptimal solution. Finally, a more generic and cohesive categorization of facial expressions is demanded for fuzzy classification of facial expressions in practical scenarios. In practice, there are not only the seven facial expressions specified above, thus forcing recognition of expressions into one of those prototypic facial expressions might not always be feasible or desirable on all occasions. Moreover, in some application scenarios, an improvement might be available if the targeted expressions can be re-categorized into groups and described in a generic dimension (perhaps on an application-related basis) rather than classifying them into one of those pre- specified categories.

The aforementioned issues have motivated us to devise a novel strategy for re-categorization of facial expressions for practical applications in this study. The topic of "how emotions are perceived by a human observer" is controversial [172], [173], [174], [175]. After many years of debate, psychologists have not yet come to an agreement. Russell [21] suggested that the illusion of emotion is rather more discrete than interconnected as a result of experimental setup and methodology adopted in psychological research, and in a self-reported psychological study, they established a two-dimensional circumplex model of emotion. Their model has enabled us to understand the relation between various emotions, and provides us with a theoretical basis for regrouping of emotions, and so do corresponding facial expressions. More specifically, in the light of two prominent dimensions in the model of emotion, novel categorizations of facial expressions for practical scenarios can be derived to facilitate the introduction of facial expression recognition into applications, such as the design of affective human-machine interaction [176], [177], artificial intelligence in education (AIED) [178], affective design for user-centred system personalization [107], and other social psychology research.

This chapter thoroughly reviews of the theoretical background of our proposed approach for categorization of facial expressions, and systematically explains our

proposed categorization scheme, and preliminarily verify the performance of our novel categorization on two multi-view facial expression databases.

5.2 Concepts of emotion

5.2.1 Theories of structural perception of emotions and facial expressions

Although, since Charles Darwin's [3] initial exploratory study of facial expression in 1872, psychologists have put great effort into understanding and depicting what emotion is, how many affective states there are, how are they perceived, and what is the relation among individual emotions, they have not yet come to an agreement on these issues. Generally, there are three fundamental but antithetical concepts of emotion, which have helped explain the interrelation among diverse emotional states. One proposition suggests that emotional states are discrete and independent, and a set of monopolar emotional states forms a complete cognitive recognition of emotion [172], [179], [180]. Alternatively, the other proposition assumes that emotions are interrelated and can be described in certain general dimensions. Theorists have proposed different configurations of multi-dimensional space to describe affective states and their position in emotional space [21], [181], [182]. The third and last important concept is appraisal theory, which suggests that before an emotion is signified, an appraisal of sequential events with respect to a person's concerns is carried out in the brain, and then based on the evaluation, a specific stimulus inside the brain and body is triggered to make affective related changes [183], [184], [185], [186], [187]. Appraisal theory helps us to understand the cognitive mechanism and origin of emotional states, but it does not assist us in describing the relation between emotional states.

All of these concepts have helped us to understand emotion. However, as our research objective is to re-cluster emotional states, which involves an analysis of the

interrelation of emotional states, discrete emotional theory apparently fails to fulfil our requirement, and appraisal theory does not enable us to distinguish and organize the emotional states. These points make multi-dimensional emotional states theory the most appropriate theoretical foundation for our research. In the following section, a thoroughly review configurations of multi-dimensional scaling of emotional states is presented.

5.2.2 Multi-dimensional space of affective states

Most of the current research on multi-dimensional space representation of emotion adopts multi-dimensional scaling [188], which is an ordination technique for revealing the correlation of individual data in a dataset through locating each data point in N-dimensional space, and factor analysis. With this technique, emotion theorists have proposed several configurations of emotion space.

Frois Wittman [189] was one of the pioneers to contribute the first studies describing the correlation among various facial expressions. Woodworth [181] introduced a scale of six steps to identify facial expression, including: (1) *love*, *happiness*, *mirth*; (2) *surprise*; (3) *fear*, *suffering*; (4) *anger*, *determination*; (5) *disgust*; (6) *contempt*; (7) *scattering*. Schlosberg [190] adapted Woodworth's categories, and observed two dimensions of emotions: pleasantness versus unpleasantness and attention versus rejection, and found that the geometric emotion model was in a circular fashion as illustrated in Figure 5.1. Following their previous work, they validate the circular 2-dimensional model of facial expression using Woodworth's scale judgement of facial expression pictures. In 1954, they introduced a third dimension, the level of activation, into their model [182] as illustrated in Figure 5.2. In self-reported research of similarity analysis of facial expression, Abelson and Sermat [191] also presented emotions in a 2-dimensional representation of facial expression, with one dimension indicating pleasant-unpleasant and the other describing tension-sleep.
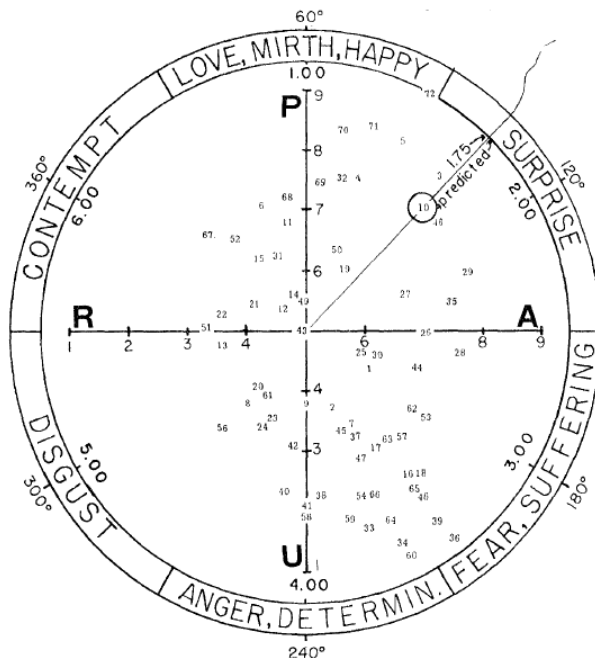
Figure 5.1: Six scales of emotions depicted in a circular fashion (taken from [181]).
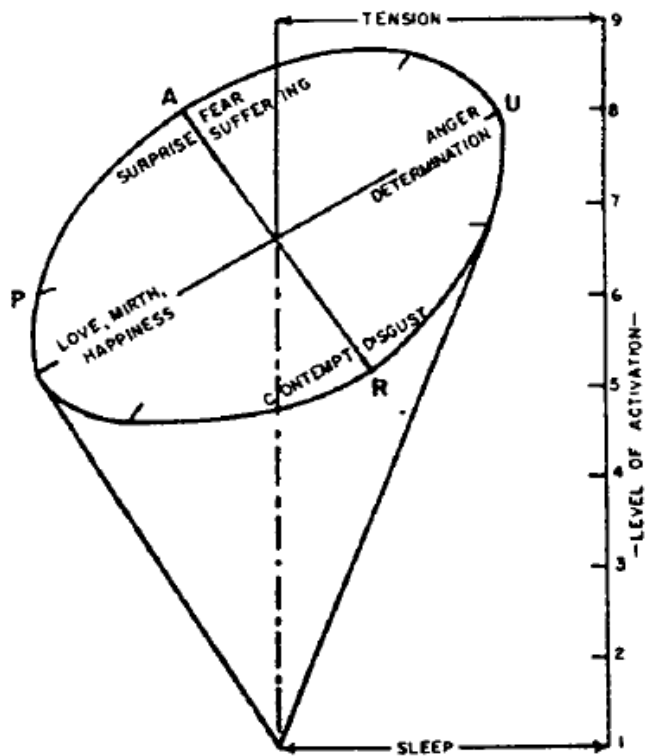
Figure 5.2: The third dimension of facial expression (taken from [182]).

Russell [8] contributes another advance in describing facial expression in general dimensions. He introduced a model of eight concepts in a circular manner in a 2-dimensional space as shown in Figure 5.3.
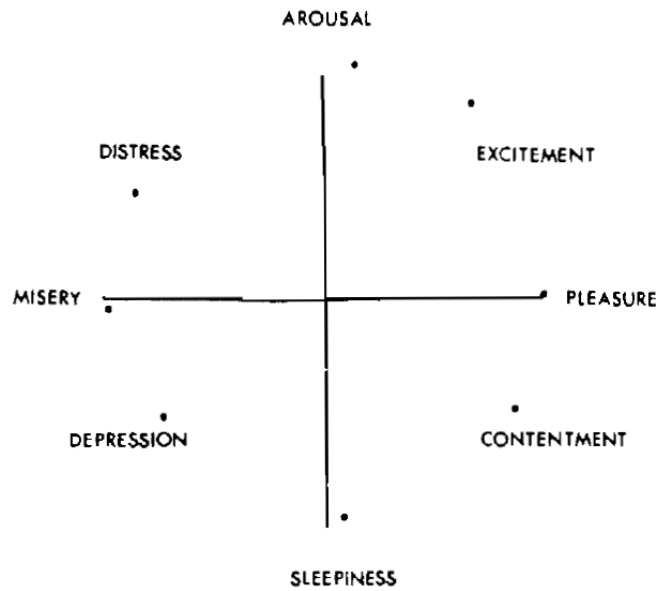
Figure 5.3: Eight affect concepts in a circular model (taken from [8]).
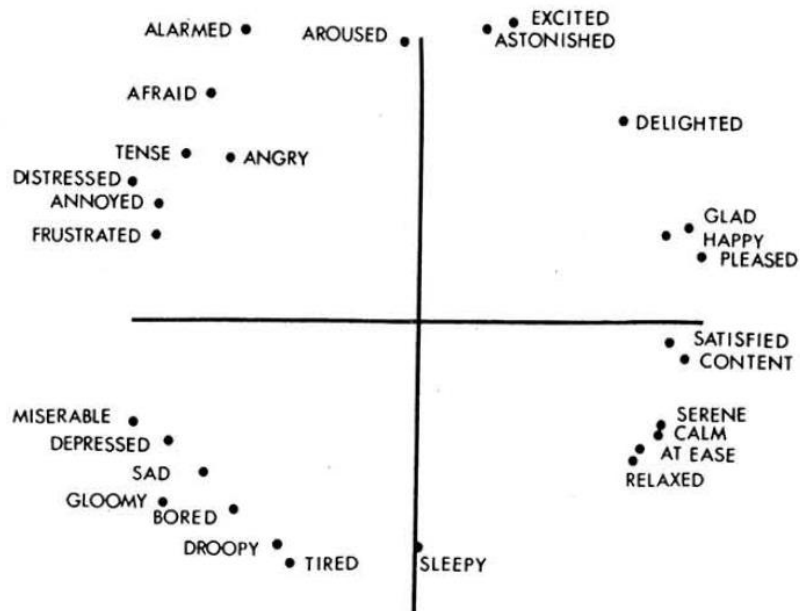
Figure 5.4: Visualized 28 affective states in the circumplex affective model (taken from [8]).

In a psychological experiment, he pinned down 28 affective states including basic emotions in the circumplex emotional axes system as shown in Figure 5.4, and he further specified that the horizontal axis indicate the intensity of pleasant to the right extreme and the intensity of "un-pleasant" to the left extreme, while the vertical axis' upper extreme represents level of arousal.
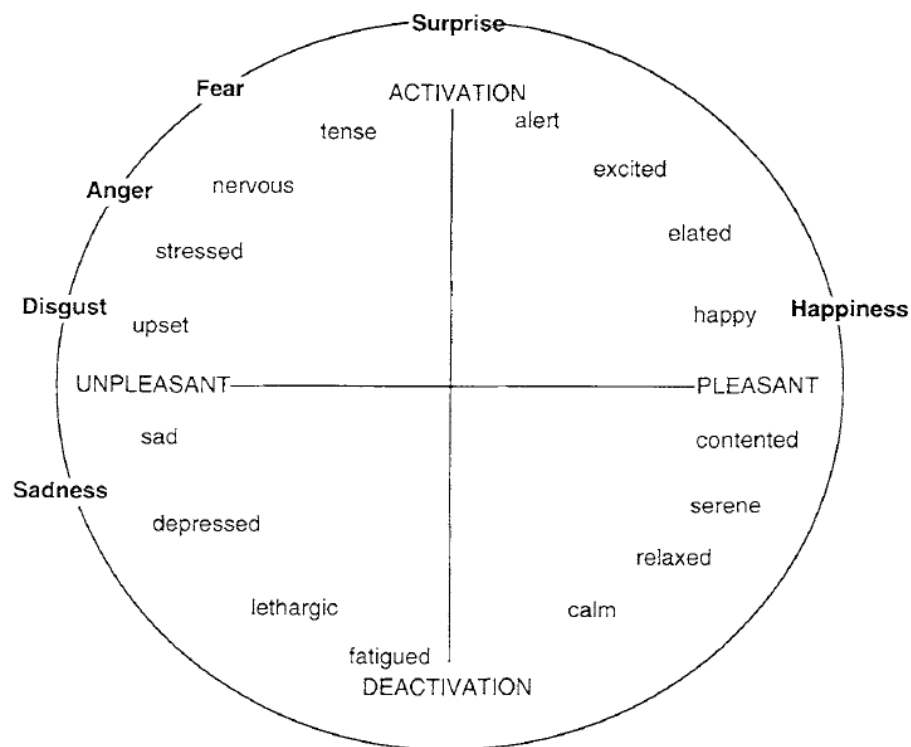
Figure 5.5: Prototypic emotional states fall in the semantic affective model suggested by Barret and Russell (taken from [21]).

Barrett and Russell [21], [192] integrate various structures of emotion, and postulate the semantic structure of core affects, which they describe in the emotional space with 2 prominent dimensions (level of pleasantness and level of activation) and a set of core affects, as shown in Figure 5.5. They revised the y-axis of the affective model as the degree of arousal, which reflects the bodily effect of emotional stimulus. In the new

configuration, the prototypic emotions are located in a circular fashion in the model with other prominent affects in the continuous pleasant and arousal orthogonal plane. This model also has suggested that emotions falling closely together in the affective model indicate a strong inter-relation and high similarity, and vice versa.

This conceptual continuous geometric emotional model has enabled us to understand the correlation and similarity of these affective states and provides a theoretical foundation for re-categorized emotional states, so as to group facial expressions further for practical applications.

5.3 Re-categorization of facial expressions for practical application

Although many automatic facial expression recognition systems have been developed and introduced, most of them attempt to classify facial expression into basic prototypic emotions. However, a recognition scheme that does not take into account the specific application scenario, would certainly restrict its performance, robustness, and validity, and even in some cases cause a system failure. For example, without considering the illumination condition of the application scene, a facial expression recognition system's robustness would be highly affected; without considering desired facial expressions to identify, a facial expression system's performance would be suboptimal compared to optimized counterparts. In this section, a variety of novel categorization schemes of facial expressions based on the circumplex model of emotion is presented to serve a broad range of applications with different purposes.

5.3.1 Grouping with respect to the degree of pleasantness (Valence)

The idea of positive and negative emotions has existed for a long time. Not only does it exert an impact on one's behaviour but also affect other individuals' attitude and performance in a group. As an influential and prominent dimension of emotion, many researchers have approached it differently but the idea behind it is the same, which is to evaluate the emotions' personal and social impact. Based on Barrett and Russell's conceptual theory of emotion [21], we further group the prototypic facial expressions with respect to their position in the circumplex affective model along the x-axis which indicates the unpleasantness-pleasantness dimension of facial expressions. Although grouping with respect to y-axis which is the dimension of activation is also a promising research direction, taking consideration of time requirement for the experiment, it is not explored in this study.

Balanced grouping: This grouping method re-categorizes emotions into three groups comprising *positive affect*, *negative affect* and *surprise*. According to the circumplex affective model [8], the y-axis separates positive and negative emotions. Positive emotions are those emotions with a projection to the positive segment on the x-axis, while negative emotions are emotions which project negatively on to the x-axis. As illustrated in Figure 5.6, the yellow area represents the negative category, and the pink area implies the positive category. It is obvious that emotions such as *upset*, *stressed*, *depressed*, *anger*, *disgust*, and *sadness* fall into the negative section of the x-axis, and *happiness*, *contented*, *relaxed*, and *excited* positions in the right section of x-axis. *Surprise,* with a projection at the origin on the x-axis, indicates that *surprise* as an emotion is neither positive nor negative, which therefore makes *surprise* a unique emotion category. And all other facial expressions that do not belong to either the positive or negative group are also categorized into the same group with *surprise*, such as *neutral*, to construct another expression category. The grey area shown in Figure 5.6 illustrates this category of other expressions. The term 'balanced' indicates that this configuration of categorization retains both a positive category and a negative

category, in contrast to a positive-only grouping and negative-only grouping. An example of an application scenario of this categorization is an automatic user experience feedback collection system that collects both positive and negative feedback.
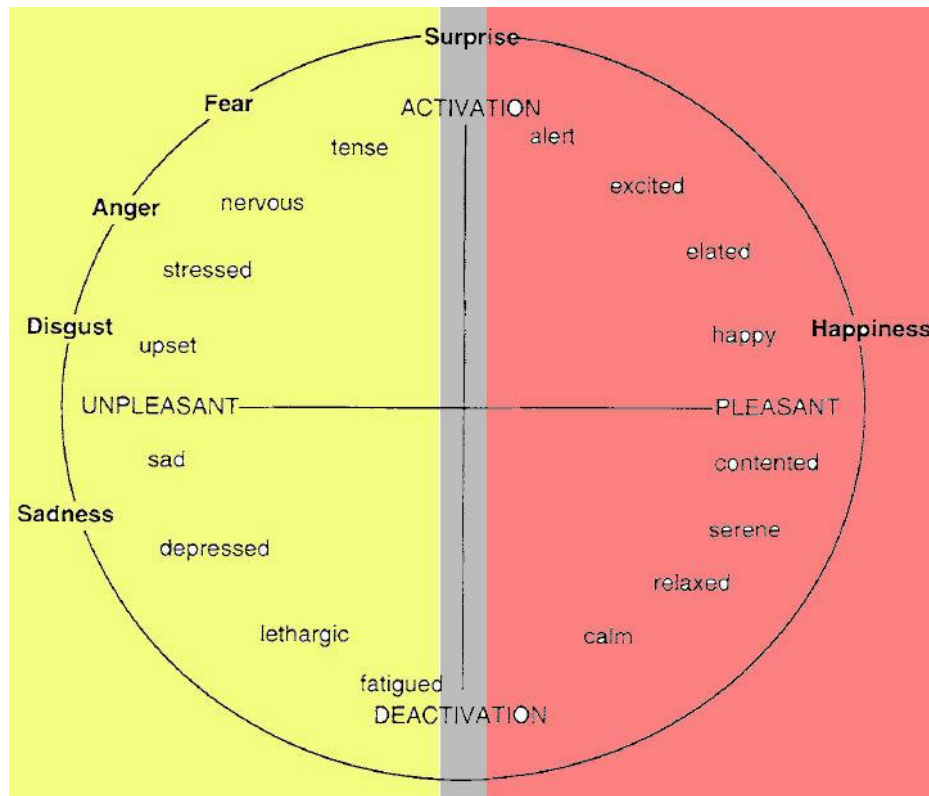
Figure 5.6: The positive affect, negative affect, and *surprise*.

Positive only grouping: this recognizes only positive emotions. Emotions are divided into two groups, positive emotions and other emotions. A facial expression recognition system that is configured with this kind of categorization only recognizes positive emotions, and all other emotions are assigned into the same group. As illustrated in Figure 5.6, only emotions from the pink area or with a projection to the positive section of the x-axis, such as *happiness*, *contented*, *relaxed*, *calm*, etc., are recognized by the system, while emotions from the grey area and yellow area are regarded as the same.

For example, a *happiness* facial expression will be recognized by the system as the positive category, and *surprise* is recognized as the "other expression" category.

Negative only grouping: in contrast to the positive grouping, a system that adopts a negative grouping categorization only recognizes negative emotions, and other emotions are, regardless of their differences, sorted into the same group. The yellow area in Figure 5.6 illustrates the negative category of this categorization, and the pink and grey areas make up the other emotion category. This configuration of categorization is suitable for adoption in practical scenarios where a user's negative emotion is of principal interest to the system or system administrator. For example, in the design of an elderly care home staff management system, negative emotions are of great interest to care home staff. Once negative emotions are detected by the system, the exact video clip can be immediately forwarded to an administrator for further inspection. In this case, the detection of the facial expression serves as initiation/alarm step of a complete staff intervention process.

5.3.2 2-dimensional bipolar analysis and grouping of emotions and facial expressions

The latest work reported by Yik, Russell and Barrett [193] has organized and summarized the previous descriptive two-dimensional affective model proposed by Russell [21], Watson & Tellegen [194], Larsen & Diener [195], and Thayer [196], and they found a consensus on the prominent dimensions of affective space. These models share the same concept that emotion space can be constructed in two dimensions, where one dimension is the degree of pleasantness and one other dimension. Although they did not adopt the same criterion to describe the second dimension in the space, the concept behind it is similar, in that they adopt the "pleasant" dimension to depict feelings of that emotion and adapt the second dimension to illustrate the consequence that the particular affect induces [26]. These chosen dimensions divide the entire emotional space into four quarters and each of them represents emotions with similar

feeling and impact, as illustrated in Figure 5.7. For example, Watson and Tellegen [194] generalize these four quarters of emotions as high negative affect, high positive affect, low positive affect and low negative affect. Russell integrated another two dimensions, the distress-relaxation dimension and excitement-depression dimension, into the model.
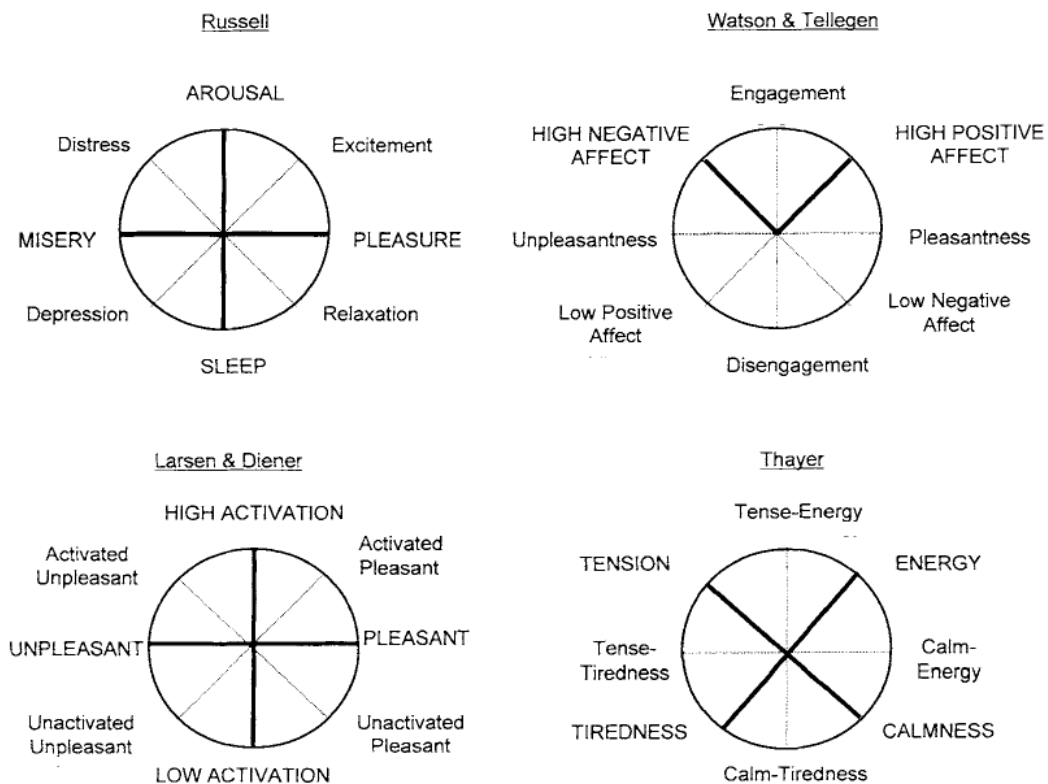
Figure 5.7: Four descriptive models of core affect summarized, [8], [21], [194], [195], [193], [196].

Quadrant grouping of emotions re-categorizes emotions with respect to both degrees of pleasantness and degree of arousal, and re-categorizes emotions into four groups or five groups including the generalised "other expression" group. Basically, emotion positions in the same quadrant of the orthogonal axis system are grouped into the same category. An illustration of quadrant grouping based on the circumplex affective

model [21] established by Barret and Russell is presented in Figure 5.8. Four areas highlighted in different colours demonstrate four categories of emotions consisting of high positive affect, low positive affect, high negative affect and low negative affect. For example, *disgust*, *anger*, *fear*, *upset* and *nervous* are grouped into high negative affect category while *excited*, *elated*, and *happiness* are sorted into high positive category. Other facial expressions including *surprise* and *neutral*, are categorized into the same group named 'other group' which constitutes the fifth group of quadrant grouping.
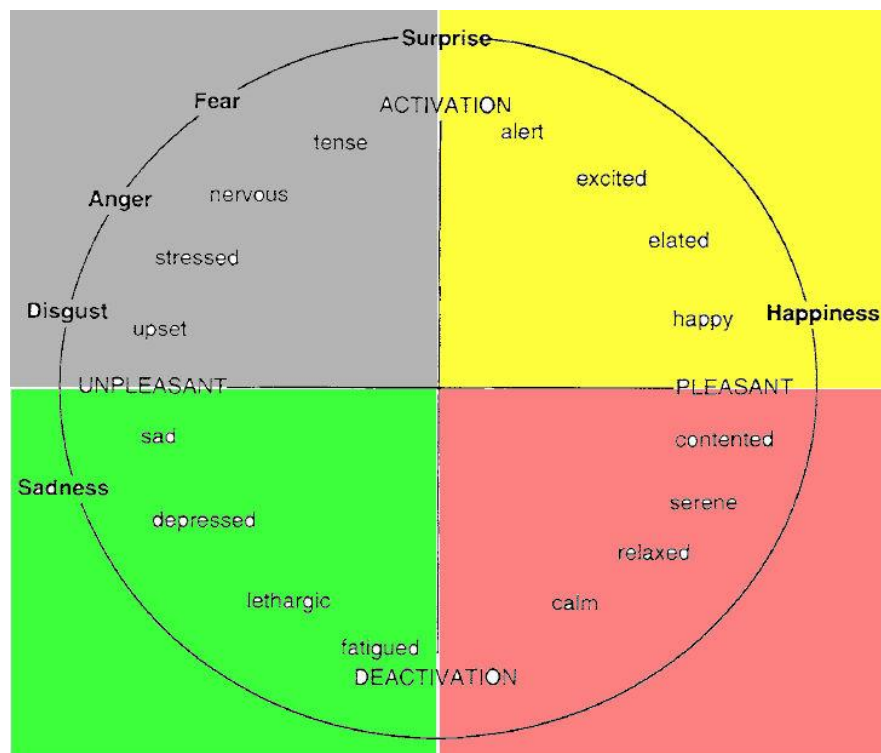
Figure 5.8: Quadrant grouping of facial expressions and emotions.

5.4 Experimental setup and results analysis:

In this section, the experimental setup of a series of preliminary experiments conducted to explore the performance and robustness of these novel configurations of

categorization of facial expressions in the application of a universal multi-view facial expression recognition is elaborated using our proposed *BBLTP* based approach with respect to a combination of viewpoint changes and variation of intensity levels. The detailed information about the *BBLTP*-based facial expression recognition system was described in Chapter 4. In this study, the local ternary pattern operator with an operating scale of 4 and tolerance threshold of 10 is utilized.

5.4.1 Data preparation and pre-processing

In order to make this experiment's results comparable with the previous experimental results, the same dataset that is simulated from *BU-3DFE* database and utilized in previous experiments is also adopted in this study. The complete information about the *BU-3DFE* database is presented in Chapter 3. Table 5.1 provides a summary of the dataset adopted in this study.

Total number of images	84,000
Tilt angles	-30º, -15º, 0º, 15º, 30º
Pan angles	-45º, -30º, -15º, 0º, 15º, 30º, 45º
Facial expressions	Six basic expressions
Intensity Levels	1, 2, 3, 4
Subjects	100

Table 5.1 Summary of the dataset generated for this study based *BU-3DFE* database.

5.4.2 Re-categorization of facial expression in practice

Since the *BU-3DFE* database only contains the prototypic facial expressions including *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*, our preliminary experiments

will re-categorize these six basic facial expressions, and evaluate the *BBLTP* based system's performance with the new categorization under different variation of views and intensity levels of facial expression.

Balanced grouping of facial expression and emotion: according to the six basic emotional states' positions in the Russell's circumplex emotional model [8], their corresponding facial expressions are re-categorized into three groups as illustrated in the following Figure 5.9. *Anger*, *disgust*, *fear* and *sadness* were grouped into the negative category while *happiness* and *surprise* were grouped into the positive category. Other expressions such as *surprise* are sorted into the "other expression" category. As a result of balanced grouping, the number of classes involved in the classification process reduces to three.
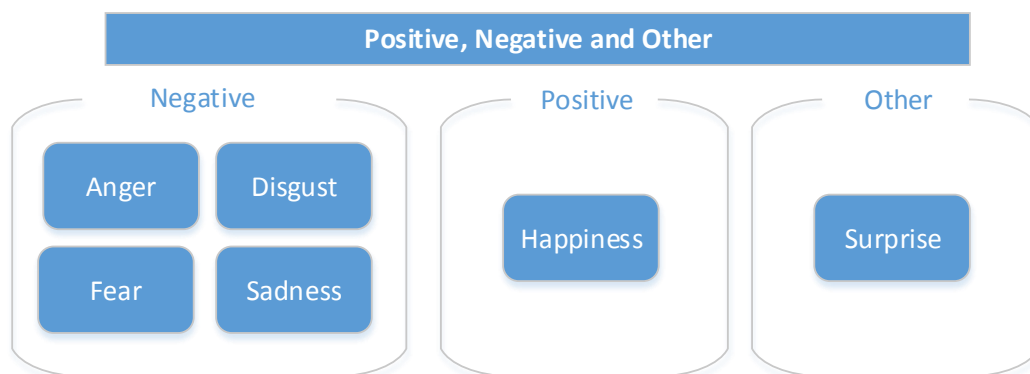
Figure 5.9: Balanced grouping of facial expressions and emotions.

Positive only grouping: re-categorizes all other facial expressions, comprising *anger*, *disgust*, *surprise*, *fear*, and *sadness*, into the same group besides *happiness*, as illustrated in Figure 5.10. Regardless of which basic emotion class the image belongs to, as long as the image does not represent *happiness*, it is assigned to the "other" category. In other words, the system recognizes the facial expression of *happiness*

only. With this configuration, only positive expressions from the application scenario will be collected.

Figure 5.10: Positive only grouping.

Negative only grouping: negative grouping of facial expression targets to recognize all negative facial expressions including *anger*, *disgust*, *fear* and *sadness*. Facial expressions other than negative facial expressions are considered to be from the same group. An illustration of the negative grouping is given in Figure 5.11. Only negative expressions in the scene are captured by the system, and all other expressions are ignored. The benefit of negative only grouping is that the complexity of classification is reduced to two.
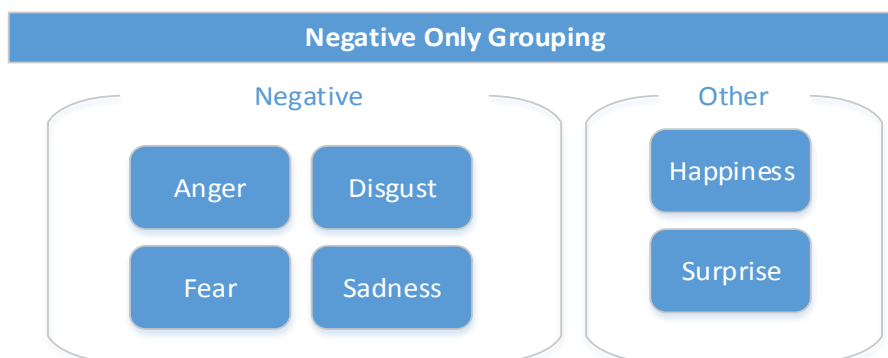
Figure 5.11: Negative only grouping

Quadrant grouping of facial expressions: in practice, six basic facial expressions are divided into four groups, including high positive group, high negative group, low negative group, and *surprise*. The low positive category is excluded from our experimental setup because none of the prototypic facial expression is from the low positive facial expression category and the majority of available facial expression databases consist of only seven prototypic facial expressions. The group of other expressions is added because *surprise* does not belong to any quarter of the orthogonal axis system. As illustrated in Figure 5.12, the facial expressions of *anger*, *fear* and *disgust* form the high negative class, and facial expressions of *happiness*, *sadness* and *surprise* form the high positive, low negative, and "other" classes respectively.
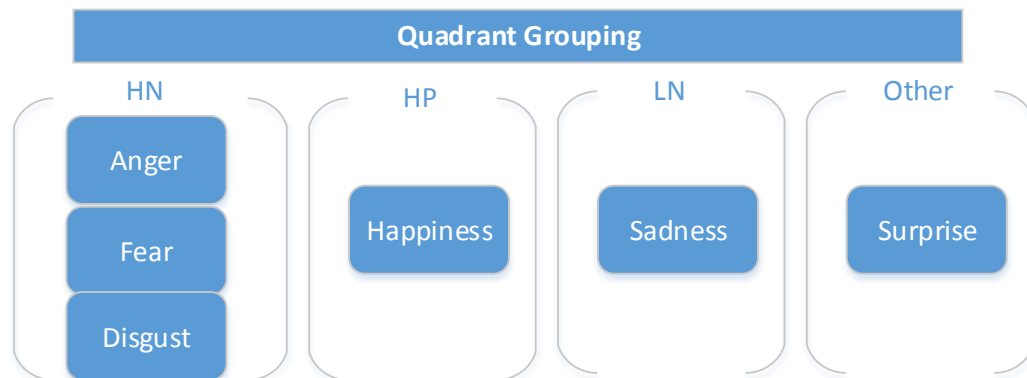
Figure 5.12: Quadrant grouping of facial expression, HN represents the high negative category, HP refers to the high positive category, and LN is the low negative category.

5.4.3 Experimental results and analysis

To evaluate the performance and robustness of these new categorizations of facial expressions, several experiments for each configuration are therefore conducted, including experiments to verify the impact which pan and tilt angle changes exert on recognition accuracy, and experiments to examine the recognition accuracy with

respect to various intensity levels of facial expressions, and in the end take into account of both factors. It should be pointed out that the classification accuracy obtained in this section should not be compared with experimental results presented in previous chapters since they adopt different categorization scheme.

5.4.3.1 Balanced grouping

The overall performance which the *BBLTP*-based system achieved was 93.81% in terms of classfication accuracy, as tabulated in Table 5.2. The experimental result has implied that the new categorization provides a good generalization within the same new categories and reasonable separation between different new categoires, and as a result, significantly reduces the overall classification difficulty of the system.

BBLTP	Overall
Accuracy	93.81%

Table 5.2: The overall system performance when adopting the balanced grouping of facial expressions.

Generally, as shown in Figure 5.13, the overall performance of the system gradually degrades with respect to the total number of intensity levels the proposed system can cope with, which is similar to the trend seen with the original prototypical facial expression categorization. The highest classification accuracy the system operates at is 96.56%, which is about 8% higher than the performance observed at the intensity level 1. Compared with the original grouping of six facial expressions, the system's operating performance at the lowest and highest levels has increased by around 13% and 10% respectively.
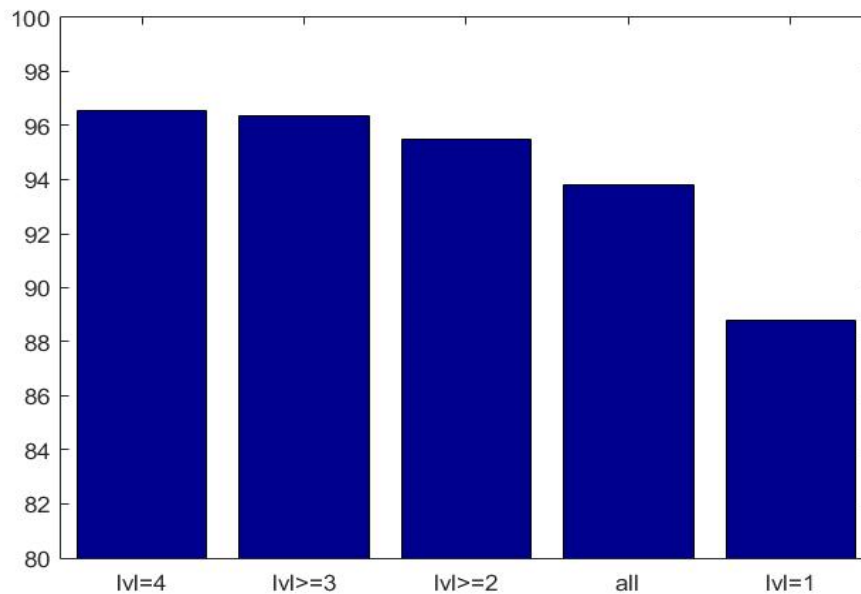
Figure 5.13: The system's operating performance with increasing numbers of intensity levels.

As shown in the classification confusion matrix, which is demonstrated in Figure 5.14, the negative facial expression group is classified with the highest accuracy while facial expressions that belong to the positive group are classified the worst with 85.57% accuracy. In order to interpret the result, the original confusion matrix of the system with the original prototypical facial expression categorization is also studied, as was shown in Figure 3.21, and it is observed that with the original categorizations, the primary factor that affects the classification accuracy of negative expression including *anger*, *disgust*, *fear* and *sadness*, is that they are to a great extent confused within each other. Therefore, sorting these negative expressions into the same category increases the system performance substantially. On the other hand, the confusion of *happiness* and *surprise* against other negative expressions accumulates to the same category, and thus the classification accuracy of these two categories decreases compared to their performance with the original categorization. The expressions of the positive category and "other" are misclassified as negative by 13.87% and 9.18% respectively.

Figure 5.14: The classification confusion matrix of the system after adopting balanced grouping.

Figure 5.15 reflects the system's operating performance at 35 different views. It is shown in the Figure that the classification accuracy when the system is operating at tilt angle of 0° and 15° is similar, and a negative tilt angle contributes a greater degradation over the classification accuracy of the system. Compared with the original performance of the system using the prototypic facial expression categorization, the balanced categorization has delivered a performance at the tilt angle of 0° and 30° with an average classification accuracy of 95.61% and 90.88%, which is 7% and 16% higher than the original system respectively. The best classification accuracy of the system is observed at tilt angle of 0° and pan angle of −30°, and the worst is observed at tilt angle of −30° and pan angle of −45°. By studying Figure 5.15 and Figure 3.22 (b), it is obvious that the overall performance at each recognition view has been substantially improved, and so does the usability of the system.
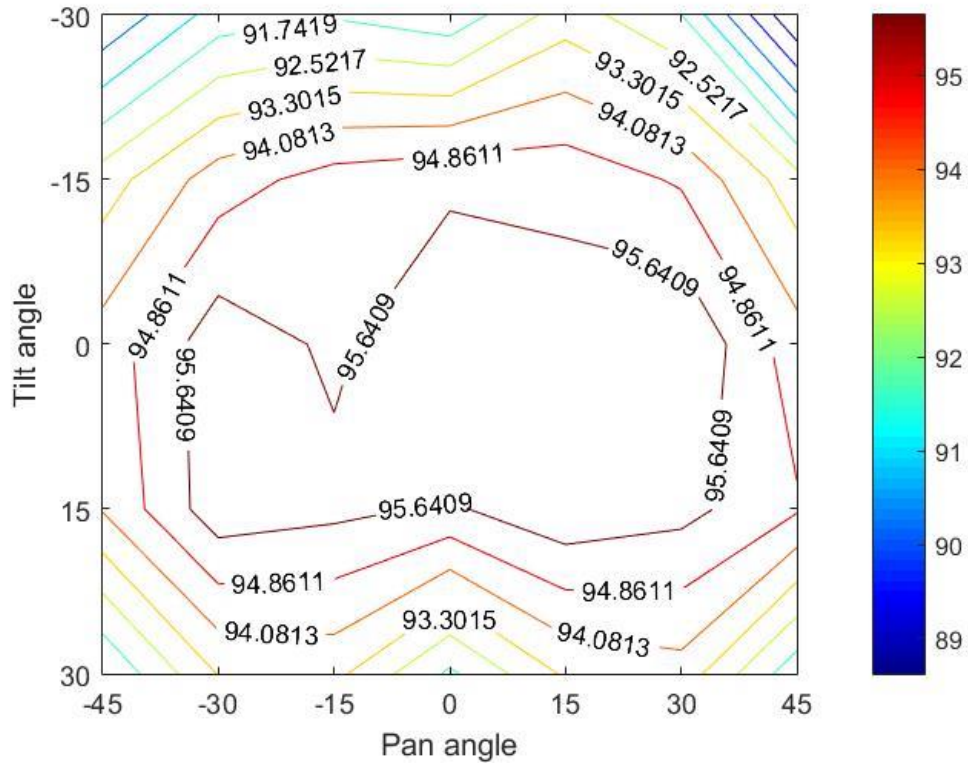
Figure 5.15: The classification accuracy of the system at different views using the balanced grouping of facial expressions.

5.4.3.2 Positive only grouping

BBLTP	Overall
Accuracy	96.15%

Table 5.3: The overall performance of the system after adopting the positive only grouping of facial expressions.

As shown in Table 5.3, the system with the positive only grouping of facial expressions yields a classification accuracy of 96.15%. This result implies that the

positive expression category is easily distinguishable from other expressions, and can be recognized by the system with high accuracy.
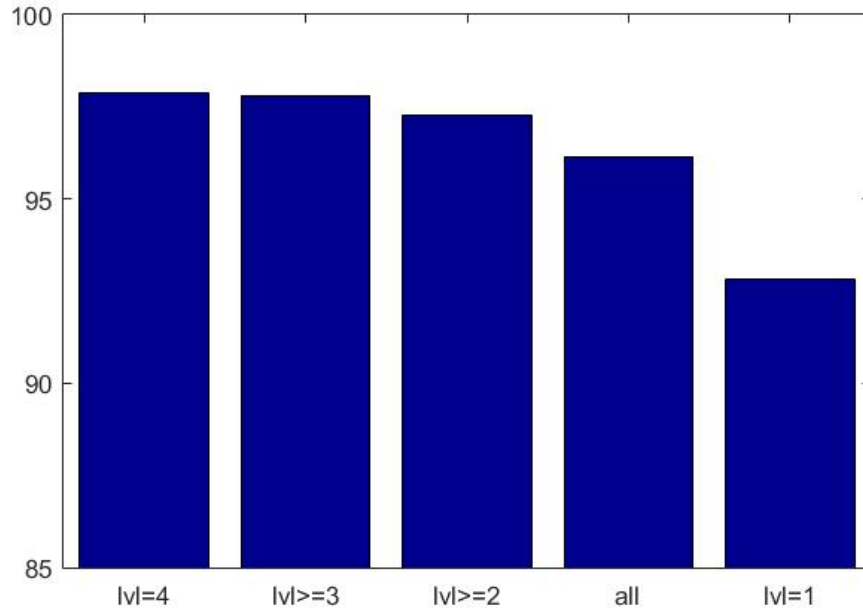
Figure 5.16: The system's operating performance with increasing numbers of intensity levels using positive only grouping.

It is demonstrated in Figure 5.16 that the classification accuracy the system is operating at decreases with respect to the number of intensity levels the system can recognize. The difference in performance at the highest and lowest intensity levels is around 5%. The recognition accuracy at the lowest intensity level is 92.83%. Also, as shown in the Figure 5.16, the overall performance of the system degrades no more than 1% at above 97% until recognition of facial expression images of the lowest intensity level are included in the test, which, again, indicates that the recognition at lowest intensity level is still the most influential factor for the overall performance of the system. By studying the classification confusion matrix of the system (as shown

in Figure 5.17), it is observed that facial expressions of the positive category are recognized less well than those of "other" category, by around 14%.

Figure 5.17: The classification confusion matrix of the system after adopting positive grouping.
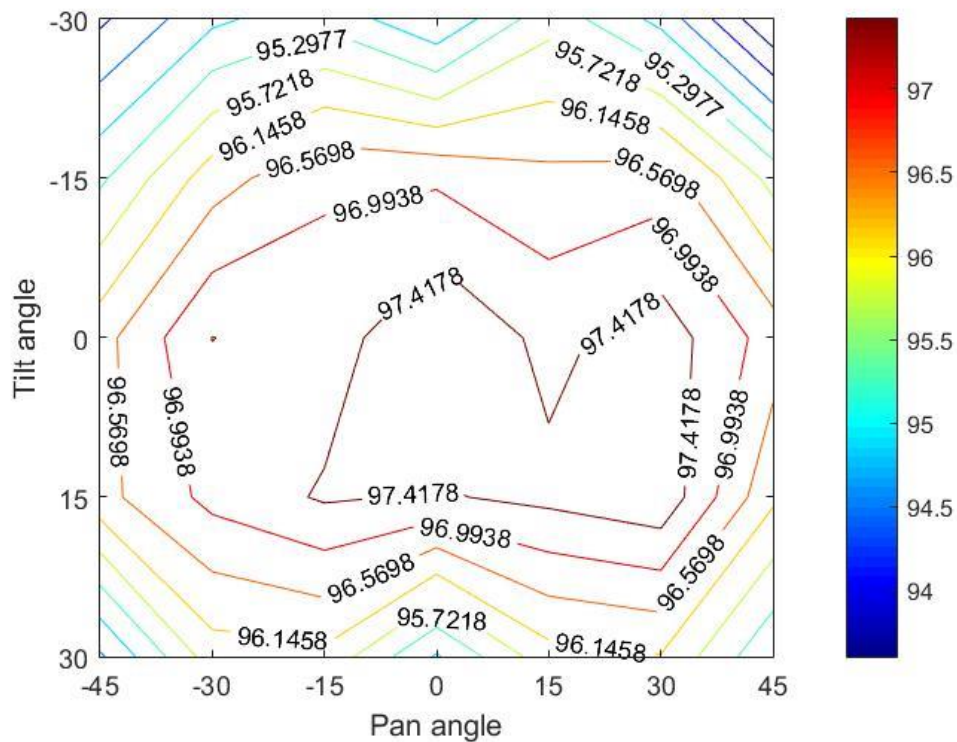
Figure 5.18: The classification accuracy of the system operating at different views using the positive only grouping of facial expressions.

After adopting the positive only grouping of facial expressions, the best performance of the system is observed at a frontal view, with classification accuracy of 97.84%, while the worst performance is again seen at tilt angle of 30° and pan angle of −45° with classification accuracy of 93.18%. The negative tilt angle remains the most significant factor to affect the system performance because the classification accuracy the system yields at −30° is about 1% lower than it produces at 30°. The general fluctuation of the system's performance with respect to the variation of views is presented in Figure 5.18.

5.4.3.3 Negative only grouping

BBLTP	Overall
Accuracy	90.38%

Table 5.4: The overall performance of the system after the negative only grouping of facial expressions is applied.

As tabulated in Table 5.4, the overall performance of the system with the negative only grouping of facial expressions is 90.38%. By inspecting the system's performance with respect to intensity levels, as illustrated in Figure 5.19, it is observed again that the system's performance is generally stable around a classification accuracy of 93% unitl recognition of facial expressions at the lowest intensity level are included. The classification accuracy at the lowest intensity level is 84.58%, which is about 8% lower than the averaged recognition rate observed without inclusion of the lowest intensity.
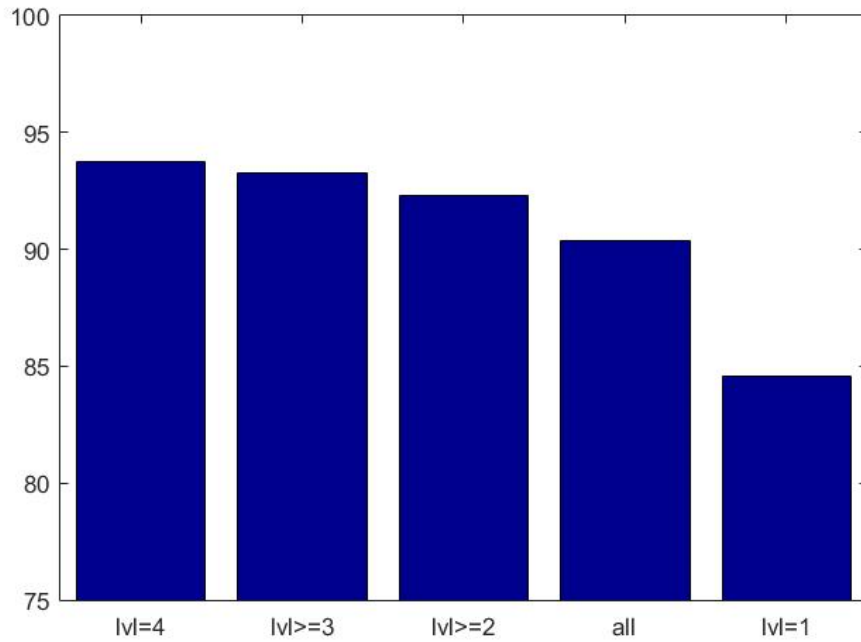
Figure 5.19: The system's operating performance with increasing numbers of intensity levels using negative only grouping.

Figure 5.20: The classification confusion matrix when the negative only grouping is applied in the univerl multi-view facial expression reocngition system.

It is shown in Figure 5.20 that facial expressions of the negative category are misclassified by 5.8% while the "other" category is misclassified as the negative

category by 17.2%. Facial expressions of the negative category are generally classified better than the "other" category.
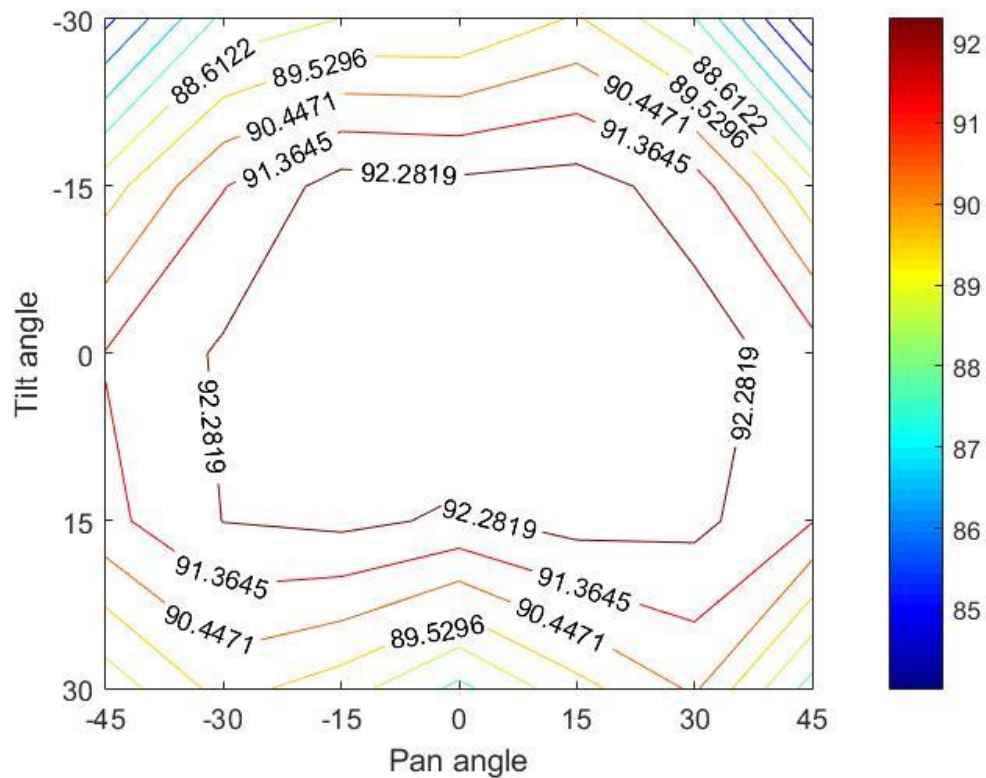
Figure 5.21: The classification accuracy of the system operating at different views using the negative only grouping of facial expressions.

By studying Figure 5.21, it can be observed that the overall performance of the system with respect to the pan angle is approximately symmetrical because the average recongtion accuracy from positive and negative sections of pan angle is nearly the same. The worst performace of the system is observed at tilt angle of $-30°$ and pan angle of $-45°$ at 83.1%, while the best performance is observed at a frontal view at 93.2%, which is about 10% higher that the worst performance view. Interestingly, the average performance of views at pan angle of $\pm15°$ is about 1% higher than that of pan angle $0°$, which suggested that pan angle of $0°$ is not the optimal for facial

expression recognition. The negative tilt angles are again observed to affect the system's performance to a greater extent than postive tilt angles.

5.4.3.4 Quadrant grouping

BBLTP	Overall
Accuracy	86.16%

Table 5.5: The overall performance of the system after the quadrant grouping of facial expressions is applied.
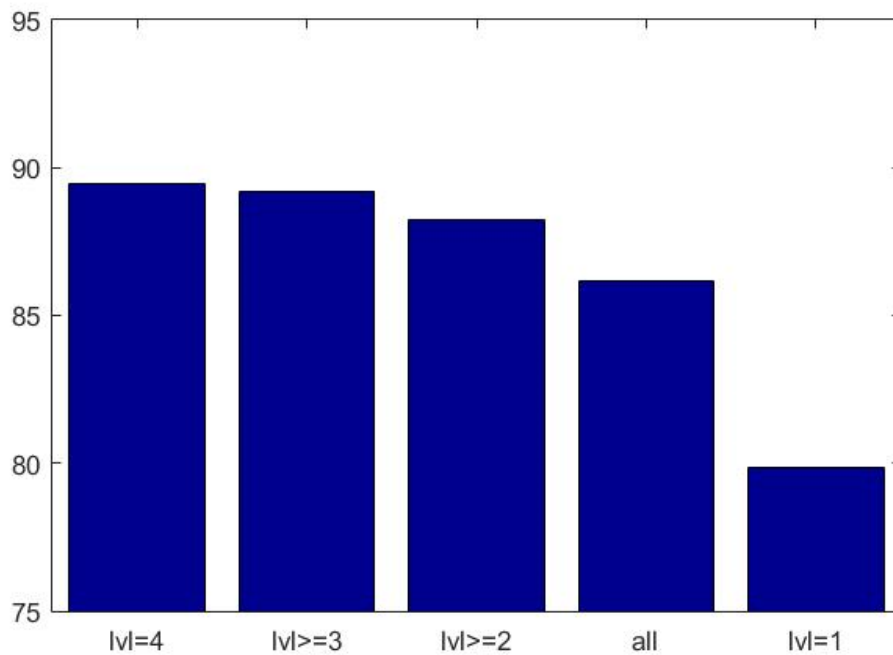
Figure 5.22: The system's operating performance with increasing numbers of intensity levels after quadrant only grouping is applied.

The overall performance of the system after adopting the quadrant grouping of facial expression as tabulated in Table 5.5, and is 86.16%. By studying the performance of the system operating at different intensity levels of facial expression, which is illustrated in Figure 5.22, it can be observed again that the system's performance drops as the intensity level of facial expression that the system is required to process lowers. It is also observed that recognition of facial expression at the lowest intensity is most difficult with a recognition accuracy of 79.87%, which is about 10% lower than the recognition rate observed at the highest intensity level alone. As a result, once recognition of facial expression at the lowest intensity is included in the testing, the overall performance was degraded by slightly over 2%.

	HN	HP	LN	Other
HN	90.531	2.564	5.060	1.845
HP	11.464	87.157	0.771	0.607
LN	30.236	1.321	66.907	1.536
Other	6.500	1.007	1.186	91.307

Figure 5.23: The classification confusion matrix when the quadrant grouping is applied. *HN*, *HP*, and *LN* represent the high negative category, high positive category, and low negative category respectively.

As shown in Figure 5.23, facial expressions of the high positive category and "other" category are recongized with over 90% accuracy while low negative expressions are classified the worst with a recognition rate of 66.9%, which is about 24% lower than the recognition rate of the "other" category. As suggested by the classificaton

confusion matrix of the originial prototypic facial expression categorization, which is shown in Figure 3.21, the separation between performance for high negative expression and low negative expression is extremely poor as *sadness* is misclassified with high negative expressions including *anger*, *disgust* and *fear* to a large extent. As a result, the misclassification rate between low negative and high negative expressions is high.
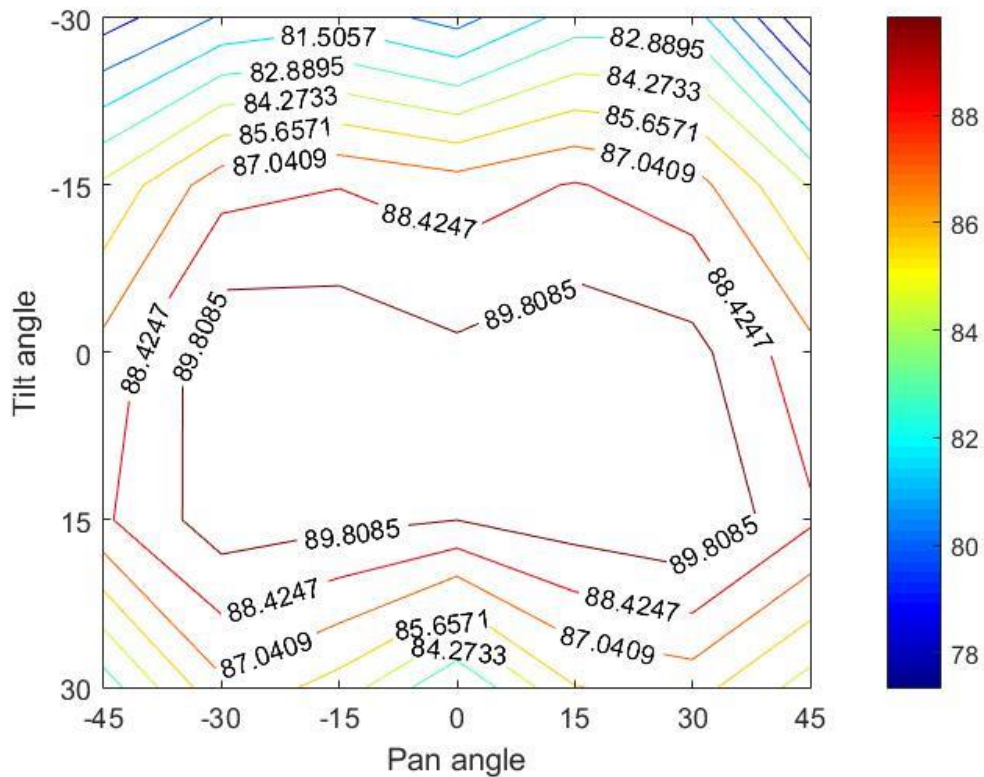


Figure 5.24: The classification accuracy of the system operating at different views using the quadrant grouping of facial expressions.

As shown in Figure 5.24, with the quadrant grouping of facial expressions, the view with the highest recognition rate is observed at tilt angle of 0° and pan angle of 30° with classification accuracy of 91.19%, while the worst view for facial expression recognition is again observed at tilt angle of −30° and pan angle of −45° with

recognition rate of 75.97%. By calcuating the average performance of the system with respect to pan angles, it is discovered that recognition rate at both pan angle of $\pm15°$ and $\pm30°$ outperforms the pan angle of 0° by 2%, and recognition from negative and positive pan angles is almost the same. By inspecting the system's performance with respect to tilt angle, a near bell-shape performance curve is seen, which peaks at tilt angle of 0° with a sharp increasing upward slope on the left climbing section and a smoother downward slope on the right, which again implies that negative tilt angle has a stronger influence over the performance of the system.

In conclusion, according to the experimental results obtained from this series of preliminary experiments of these novel categorizations of the facial expressions, it can be summarized that employing the novel categorization of facial expressions can generally improve the classification accuracy of our proposed facial expression recognition system, and for some categorizations, including the balanced grouping, positive only grouping, and negative only grouping, the overall performance of the system is significantly improved as the original confused negative expressions such as *anger*, *disgust*, *fear*, and *sadness* are sorted into the same category. The new categorizations offer a novel, generic, and broader perspective to interpret a subject's facial expressions with respect to the prominent dimensions in the affective space and extend the application of facial expression recognition into more practical application scenarios.

5.5 Conclusion

Starting from a review of the theory of emotion, in this chapter, a preliminary study of categorization of facial expressions for practical applications is presented. Based on the circumplex affective model of emotion, four different groupings of facial expression are presented, including balanced grouping, positive only grouping, negative only grouping, and quadrant grouping. In addition, a series of experiments are also carried out to evaluate the performance of the *BBLTP*-based universal multi-

view facial expression recognition system with these novel categorizations, where promising results have been observed.

The next Chapter, Chapter 6, will briefly summarize the contents of the thesis and highlight the most significant experimental results and findings observed in this complete study, and finally concludes the thesis by outlining some important possible future research directions.

Chapter 6

Summary, final conclusions, and suggestions for future work

This chapter revisits and reviews all the work that is included in this thesis, and then restates the key contributions of this thesis to the research field. By way of a conclusion to the work reported, it finally summarizes some limitations in this work and, at the same time, highlights some promising potential future research directions in relevant areas.

Section 7.1 presents a general summary of the work that has been reported in this thesis, and restates the key contributions of this thesis to the related research field. Section 7.2 outlines the potential directions for future research.

6.1 Key achievements and concluding remarks

Real time automatic recognition of facial expressions in unconstrained environments is a challenging research area due to a variety of limitations and issues introduced both by the characteristics of the subject of interest and the operational environment in unconstrained application scenarios. The problems often encountered include less pronounced facial expressions, illumination conditions, large head pose variations, occlusions, accessories inclusion, multiple subject recognition problems, impractical categorizations of facial expressions. Issues such as these have influenced achievable classification accuracy, usability, and robustness of an automatic facial expression recognition system in practical scenarios, and restrict the potential for further applications of automatic facial expression recognition technology in broader areas.

The work reported in this thesis has aimed to resolve some of these aforementioned issues that are raised when deploying an automatic facial expression recognition system in less constrained or unconstrained environments. Through an extensive series of investigations, the reported study has presented several solutions to address three fundamental issues, specifically: large pose variations, less pronounced facial expressions, and impractical categorizations of facial expressions reflecting emotion.

First, an in-house multi-view facial expression database has been designed and collected to allow us to conduct a detailed research study of the effect of out-of-plane pose angles on the performance of a multi-view facial expression recognition system. This novel database includes several factors that give it a correspondence to common "real life" data, such as accessory inclusions, light shadows on the face, and covering different ethnic groups, while it also contains the whole range of prototypic facial expressions collected from 7 pan views. This database makes possible research to conduct a preliminary investigation of the robustness and general performance of a proposed facial expression recognition system using what is, effectively, a practical dataset.

Second, in order to address the important issues, which arise, especially the large head pose variations presented in practical scenarios and less pronounced facial expressions in spontaneous facial expressions, a thorough investigation of an appropriate texture feature representation is conducted. Through a series of experiments under a new proposed system framework for multi-view facial expression recognition, we have observed the potential application of the local ternary pattern in describing facial expressions with large head pose variations. Through extending the original local ternary pattern to extract features from any radius, it is revealed that the local ternary pattern descriptor can be deployed effectively and efficiently to describe facial expressions acquired from a range of views and under a variety of intensities of facial expressions. With an empirically selected scale and threshold of a block based local ternary pattern operator, we have devised a multi-view facial expression recognition system which significantly outperforms state-of-the-art facial expression recognition systems across three different databases, including the *BU-3DFE* database, the *CK+* database, and the *JAFFE* database. The proposed system was also tested on the specially acquired in-house multi-view facial expression recognition database, which is a more practical database with various accessory inclusions, where excellent performance is also reported. To sum up, the proposed system has achieved pose invariant facial expression recognition at 35 different views ranging from $-45°$ to $45°$ pan views and $-30°$ to $30°$ tilt angles under 4 different intensities of facial expressions, including the least pronounced facial expression which are found in the *BU-3DFE* database.

Third, through a thorough investigation of the fusion of various texture features in order to seek a better feature representation for multi-view facial expression recognition system, a deficiency associated with some texture features is revealed. Specifically, it has been shown that some texture descriptors are weak in describing the general intensities of the image, such as the local binary pattern operator and the histogram of oriented gradient operator as discovered in this study. Therefore, the level of difference descriptor is introduced as a supplement for this category of texture descriptors to construct a better feature representation for a multi-view facial

expression recognition system under the proposed system framework. Through a series of experiments, it has been observed that the level of difference descriptor can significantly improve the performance of the local binary pattern and histogram of oriented gradient based feature representations. Especially, state-of-the-art performance has been achieved using the combined feature of the local binary pattern and the level of difference feature-based multi-view facial expression recognition system, which significantly outperforms the combined feature of a local Gabor binary pattern and the local binary pattern-based system with a similar setup on the *BU-3DFE* database in terms of classification accuracy, with a much more compact feature representation. In fact, the aforementioned system used a feature dimensionality of 181,248 features. The proposed system reduces this by about a factor of 72.

Fourth, to provide a more feasible facial expression categorization for practical applications, a preliminary exploration of novel categorizations of facial expression in the prominent dimensions of affective space is presented in this thesis. Adopting the spatial theory of facial expression of emotions, a range of novel categorization schemes of facial expression of emotions are presented to fulfil different application-oriented requirements and facilitate the design of an automatic facial expression recognition system in practical application scenarios. These novel categorizations provide a cohesive and broader perspective of a subject's emotional state. Implementation of some preliminary experiments using the proposed novel categorizations in the proposed system framework using the block based local ternary patterns has shown that these categorizations can offer a significant improvement on the performance of a multi-view facial expression recognition system.

In summary, in this thesis, a number of important issues relating to the practical application of facial expression recognition have been addressed, including large head pose variations and less pronounced facial expressions using novel and combined texture feature representations under the proposed system framework. In addition, a range of novel categorization schemes has been proposed to offer a broader perspective on the application of such a technology in practical scenarios.

This chapter has briefly summarised again the research problems which the study reported in this thesis has aimed to resolve, and has then revisited the primary achievements accomplished within the study described. Finally, it has highlighted the key contributions of this study. Following the general overview of the thesis, a brief analysis of the limitations of the work reported in this thesis and an outline of research directions for future studies in some important relevant search areas have been presented by way of a conclusion.

6.2 Future work

Although this thesis has described several solutions, which can be effective in resolving these issues, arising from less constrained or unconstrained environments, there still are some challenges and problems that need to be thoroughly investigated further in the future. This leads to the observation that the following two specific limitation of the work should be address as a high priority in the future.

- First, although, in this thesis, an in-house multi-view facial expression database with 7 views and 6 prototypic facial expressions has been presented, comparing this with the more comprehensive and more widely adopted databases described, it cannot offer the same amount of data as those synthesized facial expressions dataset reconstructing from a 3D facial database. Therefore, a more comprehensive and thorough facial expression database would be valuable for both researching and evaluating facial expression recognition system in real life application scenarios.

- Second, more practical categorization schemes of facial expressions are required for applications of facial expression in a wider variety of practical scenarios. In this thesis, a range of categorization schemes have been presented

to provide a broader perspective for analyzing and applying facial expression recognition technology in practical scenarios, but comparing to the numbers of potential application scenarios, these categorization schemes are far less than sufficient to meet all the requirements for emotional experience and behavioral-related research studies.

References:

[1] C. Bell, and S. Alexander, *The anatomy and philosophy of expression as connected with the fine arts*, George Bell & Sons, 1904.

[2] G. B. Duchenne, and R. A. Cuthbertson, *The mechanism of human facial expression*, Cambridge university press, 1990.

[3] C. Darwin, P. Ekman, and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[4] C. E. Izard, *Patterns of emotions*, New York: Academic Press, 1972.

[5] S. Tomkins, "Affect, imagery, consciousness. Vol. 2: The negative affects," 1963.

[6] P. Ekman and W. V Friesen, "Facial Action Coding System (FACS): A technique for the measuretnent of facial action," Palo Alto, CA: Consulting Psychologists Press, 1978.

[7] W. Friesen and P. Ekman, "Emfacs-7: emotional facial action coding system," *Unpubl. manuscript, Univ. Calif. San Fr.*, 1983.

[8] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6. pp. 1161–1178, 1980.

[9] P. Ekman, "Basic emotions," *Cognition*, vol. 98, no. 1992. pp. 45–60, 1999.

[10] P. Ekman and W. V Friesen, "Constants across cultures in the face and emotion," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[11] Y. L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*, 2005, pp. 247–275.

[12] P. Ekman, "Facial expression and emotion," *Am. Psychol.*, vol. 48, no. 4, pp. 384–392, 1992.

[13]   T. Scheff, "What are emotions? A physical theory.," *Rev. Gen. Psychol.*, vol. 19, no. 4, pp. 458–464, 2015.

[14]   P. Ekman, W. Friesen, and S. Tomkins, "Facial affect scoring technique: A first validity study," *Semiotica*, vol. 3, pp. 37–58, 1971.

[15]   C. E. Izard, "The maximally discriminative facial movement coding system (MAX)," *Unpubl. manuscript, Available from Instr. Resour. Centre, Univ. Delaware*, 1979.

[16]   P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford: Oxford University Press, 1997.

[17]   K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge: Cambridge University Press, 1982.

[18]   M. F. Valstar, *Timing is everything A spatio-temporal approach to the analysis of facial actions*. Imperial College London, 2008.

[19]   Y.-L. T. Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *Proc. IEEE Conf. Comput. Vis. Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, vol. 23, no. 2, pp. 97–115, 2001.

[20]   P. Ekman, *Emotion in the Human Face*. Cambridge: Cambridge University Press, 1982.

[21]   J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant.," *J. Pers. Soc. Psychol.*, vol. 76, no. 5, pp. 805–819, 1999.

[22]   M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," *Int. Jt. Conf. Pattern Recognit.*, 1978.

[23]   A. R. Chadha, P. P. Vaidya, and M. M. Roja, "Face recognition using discrete

cosine transform for global and local features," *2011 Int. Conf. Recent Adv. Electr. Electron. Control Eng.*, pp. 502–505, 2011.

[24] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1–14, Jun. 2007.

[25] L. Xu and P. Mordohai, "Automatic Facial Expression Recognition using Bags of Motion Words," *Procdings Br. Mach. Vis. Conf. 2010*, pp. 13.1–13.13, 2010.

[26] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," *2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG 2011*, pp. 336–342, 2011.

[27] I. I. a. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 757–763, 1997.

[28] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 42, no. 1, pp. 28–43, 2012.

[29] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, no. 02, pp. 253–263, 1999.

[30] I. Kotsia, S. Zafeiriou, N. Nikolaidis, and I. Pitas, "Texture and shape information fusion for facial action unit recognition," *Proc. 1st Int. Conf. Adv. Comput. Interact. ACHI 2008*, vol. 41, pp. 77–82, 2008.

[31] Y. Tian, T. Kanade, and J. F. Cohn, "Facial Expression Analysis," in *Handbook of Face Recognition*, New York: Springer, 2005, pp. 247–275.

[32] P. Viola, and M. Jones. "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition, 2001. CVPR 2001.*

Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 1. IEEE, 2001.

[33]   P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2001.

[34]   J. Jianxin Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, "Fast Asymmetric Learning for Cascade Face Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 369–382, Mar. 2008.

[35]   M.-T. Pham and T.-J. Cham, "Fast training and selection of Haar features using statistics in boosting-based face detection," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–7.

[36]   J. Jianguo Li, T. Tao Wang, and Y. Yimin Zhang, "Face detection using SURF cascade," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 2183–2190.

[37]   H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.*, Vol. 110, No. 3, pp. 346–359, 2008.

[38]   B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, Jul. 1997.

[39]   H. A. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, pp. 23–38, 1998.

[40]   D. Roth, M.-H. Yang, and N. Ahuja, "A SNoW-Based Face Detector," in *Proceedings of Neural Information Processing Systems*, 2000, pp. 855–861.

[41]   R. Feraud, O. Bernier, J. E. Viallet, and M. Collobert, "A fast and accurate face detector for indexation of face images," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 77–82.

[42] H. Schneiderman and T. Kanade, "A statistical method for 3D object detection applied to faces and cars," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, 2000, vol. 1, pp. 746–751.

[43] C. Huang, H. Ai, Y. Li, and S. Lao, "High-Performance Rotation Invariant Multiview Face Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 671–686, Apr. 2007.

[44] S. Parupati, R. Bakkannagari, S. Sankar, and V. Kulathumani, "Collaborative acquisition of multi-view face images in real-time using a wireless camera network," in *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, 2011, pp. 1–6.

[45] S. Liao, A. K. Jain, and S. Z. Li, "A Fast and Accurate Unconstrained Face Detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 211–223, Feb. 2016.

[46] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, 2004, pp. 91–96.

[47] M. Minyoung Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[48] Y. Mao, H. Li, and Z. Yin, "Who missed the class? Unifying multi-face detection, tracking and recognition in videos," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 2014, pp. 1–6.

[49] C. J. Pereira Passarinho, E. Ottoni Teatini Salles, and M. Sarcinelli Filho, "Face Tracking in Unconstrained Color Videos with the Recovery of the Location of Lost Faces," *IEEE Lat. Am. Trans.*, vol. 13, no. 1, pp. 307–314, Jan. 2015.

[50] W. Zhang, Y. Yu Qiao, C. Chunjing Xu, and S. Shifeng Chen, "Robust non-

rigid 3D tracking for face recognition in real-world videos," in *2011 IEEE International Conference on Information and Automation*, 2011, pp. 902–907.

[51] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, Mar. 1975.

[52] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis, and S. Kollias, "Parameterized facial expression synthesis based on MPEG-4," *EURASIP J. Appl. Signal Processing*, vol. 2002, no. 10, pp. 1021–1038, 2002.

[53] Ahlberg, Jörgen. "Candide-3-an updated parameterised face," unpublished.

[54] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and Support Vector Machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–87, 2007.

[55] F. Tsalakanidou and S. Malassiotis, "Real-time 2D+3D facial action and expression recognition," *Pattern Recognit.*, vol. 43, no. 5, pp. 1763–1775, 2010.

[56] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, *Active Shape Models-Their Training and Application*, vol. 61, no. 1. 1995, pp. 38–59.

[57] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 1998.

[58] T. Cootes and C. Taylor, "Statistical models of appearance for medical image analysis and computer vision," *Med. Imaging 2001*, pp. 236–248, 2001.

[59] T. F. Cootes and C. J. Taylor, "Statistical Models of Appearance for Computer Vision," *Direct*, vol. M, no. 1, pp. 1–124, 2004.

[60] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2D+3D active appearance models," *Proc. 2004 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR 2004)*, pp. 535–542, 2004.

[61] H. S. Lee and D. Daijin Kim, "Tensor-Based AAM with Continuous Variation

Estimation: Application to Variation-Robust Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 1102–1116, Jun. 2009.

[62] S. V. Duhn, L. Yin, M. J. Ko, and T. Hung, "Multiple-View Face Tracking For Modeling and Analysis Based On Non-Cooperative Video Imagery," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[63] R. Gross, I. Matthews, and S. Baker, "Constructing and Fitting Active Appearance Models With Occlusion," *2004 Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 72–72, 2004.

[64] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image Vis. Comput. J.*, vol. 18, no. 11, pp. 881–905, 2000.

[65] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," *Proc. Third IEEE Int. Conf. Autom. Face Gesture Recognit.*, pp. 454–459, 1998.

[66] O. Rudovic, I. Patras, and M. Pantic, "Coupled Gaussian process regression for pose-invariant facial expression recognition," *Comput. Vision–ECCV 2010*, vol. 35, no. 6, pp. 350–363, 2010.

[67] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines.," *Sensors (Basel).*, vol. 13, pp. 7714–34, 2013.

[68] T. S. Tai Sing Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959–971, 1996.

[69] M. Lyons, S. Akamatsu, et al. "Coding facial expressions with gabor wavelets."*Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998.

[70] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1, pp. 37–52, 1987.

[71] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vision Res.*, vol. 41, no. 9, pp. 1179–1208, 2001.

[72] M. Li, and B. Yuan. "2D-LDA: A statistical linear discriminant analysis for image matrix." *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527-532, 2005.

[73] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[74] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[75] N. Dalal, and B. Triggs. "Histograms of oriented gradients for human detection." *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).* Vol. 1. IEEE, 2005.

[76] Hesse, Nikolas, et al. "Multi-view facial expression recognition using local appearance features." *Pattern Recognition (ICPR), 2012 21st International Conference on.* IEEE, 2012.

[77] T. Ojala, M. Pietikainen, T. Maenpaa, M. Pietik??inen, T. M??enp????, M. Gray-scale, T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[78] P. W. W. McOwan, C. Shan, S. Gong, P. W. W. McOwan, Caifeng Shan, Shaogang Gong, P. W. W. McOwan, C. Shan, S. Gong, and P. W. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, 2005, vol. 2, pp. 914–917.

[79] L. Nanni, S. Brahnam, and A. Lumini, "A local approach based on a Local Binary Patterns variant texture descriptor for classifying pain states," *Expert

Syst. Appl., vol. 37, no. 12, pp. 7888–7894, 2010.

[80] S. Liao, W. Fan, A. S. Chung, and D. Yeung, "Facial Expression Recognition using Advanced Local Binary Patterns, Tsallis Entropies and Global Appearance Features," in *2006 International Conference on Image Processing*, 2006, pp. 665–668.

[81] C. Tsallis, "Nonextensive statistics: theoretical, experimental and computational evidences and connections," *Brazilian J. Phys.*, vol. 29, no. 1, pp. 1–35, Mar. 1999.

[82] H. Deng, L. Jin, L. Zhen, and J. Huang, "A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA," *Int. J. Inf. Technol.*, no. 303, pp. 86–96, 2005.

[83] K. Yu, Z. Wang, L. Zhuo, J. Wang, Z. Chi, and D. Feng, "Learning realistic facial expressions from web images," *Pattern Recognit.*, vol. 46, no. 8, pp. 2144–2155, 2013.

[84] L. Lying Ma, Y. Yegui Xiao, K. Khorasani, and R. K. Ward, "A new facial expression recognition technique using 2-D DCT and K-means algorithm," in *2004 International Conference on Image Processing, 2004. ICIP '04.*, 2004, vol. 2, pp. 1269–1272.

[85] O. Rudovic, "*Machine Learning Techniques for Automated Analysis of Facial Expressions,* " Ph.D. dissertation, Imperial College London, 2013.

[86] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De La Torre, "Continuous AU intensity estimation using localized, sparse facial feature space," *2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013*, 2013.

[87] Orozco, Javier, et al. "Spatio-temporal reasoning for reliable facial expression interpretation," unpublished.

[88] S. Eleftheriadis, O. Rudovic, and M. Pantic. "Shared gaussian process latent variable model for multi-view facial expression recognition."*International*

Symposium on Visual Computing. Springer Berlin Heidelberg, 2013.

[89]   O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "Expressive maps for
3D facial expression recognition," *Proc. IEEE Int. Conf. Comput. Vis. Work.*,
pp. 1270–1275, 2011.

[90]   J. Wang and L. Yin, "Static topographic modeling for facial expression
recognition and analysis," *Comput. Vis. Image Underst.*, vol. 108, no. 1–2, pp.
19–34, 2007.

[91]   U. Prabhu, J. Heo, and M. Savvides, "Unconstrained Pose Invariant Face
Recognition Using 3D Generic Elastic Models.," *IEEE Trans. Pattern Anal.
Mach. Intell.*, vol. 33, no. 10, pp. 1952–1961, 2011.

[92]   T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3D/4D facial
expression analysis: An advanced annotated face model approach," *Image Vis.
Comput.*, vol. 30, no. 10, pp. 738–749, 2012.

[93]   A. Maalej, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti, "Local 3D
Shape Analysis for Facial Expression Recognition," in *Proc. 20th Int. Conf. on
Pattern Recognition*, 2010, pp. 4129–4132.

[94]   H. Li, J. M. Morvan, and L. Chen, "3D facial expression recognition based on
histograms of surface differential quantities," *Lect. Notes Comput. Sci.
(including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol.
6915 LNCS, pp. 483–494, 2011.

[95]   J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based
on primitive surface feature distribution," *Proc. IEEE Comput. Soc. Conf.
Comput. Vis. Pattern Recognit.*, vol. 2, pp. 1399–1406, 2006.

[96]   Z. Ambadar, J. W. Schooler, and J. F. Cohn, "Deciphering the enigmatic face
the importance of facial dynamics in interpreting subtle facial
expressions," *Psychological science*, vol. 16, no. 5, pp. 403-410, 2005.

[97]   J. F. Cohn, and K. L. Schmidt, "The timing of facial motion in posed and

spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 02, pp. 121-132, 2004.

[98] I. Cohen, N. Sebe, A. Garg, L. S. Chen, T. S. Huang, L. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Underst.*, vol. 91, no. 1–2, pp. 160–187, 2003.

[99] D. Li, X. Wang, and Y. Tian, "Dynamic Facial Expression Feature Extraction and Classification Based on Candide-3 Face Model," in *2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control*, 2014, pp. 877–882.

[100] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," *Proc. - 2013 Hum. Assoc. Conf. Affect. Comput. Intell. Interact. ACII 2013*, pp. 356–361, 2013.

[101] M. Taini, G. Zhao, S. Z. Li, and M. Pietikäinen, "Facial Expression Recognition from Near-Infrared Video Sequences," pp. 1–4, 2008.

[102] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," *3rd Int. Conf. Imaging Crime Detect. Prev. (ICDP 2009)*, pp. 16, 2009.

[103] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699–714, 2005.

[104] P. Wang, F. Barrett, E. Martin, M. Milanova, R. E. Gur, C. Ruben, C. Kohler, and R. Verma, "Automated Video Based Facial Expression Analysis of Neuropsychiatric Disorders," *J Neurosci Methods*, vol. 15, no. 168, pp. 224–238, 2008.

[105] "JIBO Robot." [Online]. Available: https://www.jibo.com/. [Accessed: 03-

May-2016].

[106] A. Mehrabian, "Communication without Words," in *Communication Theory*, Second Edi., C. D. Mortensen, Ed. New jersey & London: Transaction Publishers, 2008, pp. 149–159.

[107] F. Zhou, Y. Ji, and R. J. Jiao, "Affective and cognitive design for mass personalization: status and prospect," *J. Intell. Manuf.*, vol. 24, no. 5, pp. 1047–1069, Oct. 2013.

[108] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[109] Y.-L. Tian, T. Kanade, and J. F. Cohn, *Handbook of Face Recognition*. London: Springer, 2005.

[110] F. Y. Shih, C.-F. Chuang, and P. S. P. Wang, "PERFORMANCE COMPARISONS OF FACIAL EXPRESSION RECOGNITION IN JAFFE DATABASE," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, no. 3, pp. 445–459, 2008.

[111] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random Gabor based templates for facial expression recognition in images with facial occlusion," *Neurocomputing*, vol. 145, pp. 451–464, 2014.

[112] L. Zhang and D. Tjondronegoro, "Facial Expression Recognition Using Facial Movement Features," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 219–229, Oct. 2011.

[113] R. El-Sayed, a El Kholy, and M. El-Nahas, "Robust Facial Expression Recognition via Sparse Representation and Multiple Gabor filters," *Int. J.*, vol. 4, no. 3, pp. 82–87, 2013.

[114] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kande dataset (CK+): A complete facial expression dataset for action unit and emotionspecified expression," *IEEE Conf. Comput. Vis. Pattern*

Recognit. Work., pp. 94–101, July 2010.

[115] T. Kanade, J. F. Cohn, and Yingli Tian, "Comprehensive database for facial expression analysis," *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, pp. 46–53, 2000.

[116] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006, pp. 211–216.

[117] "3dMD Face System." [Online]. Available: http://www.3dmd.com/, [Accessed: 03-May-2016].

[118] The MathWorks Inc., "Simulink 3D Animation Toolbox," *2013*. [Online]. Available: http://uk.mathworks.com/products/3d-animation/model-examples.html, [Accessed: 03-May-2016].

[119] M. Pantic, et al., "Web-Based Database for Facial Expression Analysis," *IEEE int. conference on multimedia. and expo.*, 2005, pp. 0–4.

[120] P. Lucey, et al., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops.* IEEE, 2010.

[121] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[122] D. J. Kroon, "Viola Jones Object Dection," 2010. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/29437-viola-jones-object-detection, [Accessed: 03-May-2016].

[123] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, 1992, pp. 144–152.

[124] Z. Niu and X. Qiu, "Facial Expression Recognition based on weighted principal component analysis and support vector machines," in *Advanced Computer Theory and Engineering (ICACTE)*, 2010, pp. 174–178.

[125] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

[126] K. Crammer, "On the Learnability and Design of Output Codes for Multiclass Problems," *Mach. Learn.*, vol. 47, pp. 201–233, 2002.

[127] R. E. Fan, et al., "LIBLINEAR: A Library for Large Linear Classification," *J. Mach. Learn.*, vol. 9, no. 2008, pp. 1871–1874, 2008.

[128] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." *IJCAI*. Vol. 14. No. 2. 1995.

[129] M. Pietikäinen, A. Hadid, G. Zhao, Ahonen, and T. Ahonen, *Computer Vision Using Local Binary Patterns*, vol. 40, no. 11. London: Springer, 2011.

[130] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing 2005*, 2005, pp. 370.

[131] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[132] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[133] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "A novel approach to expression recognition from non-frontal face images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 1901–1908.

[134] U. Tariq, J. Yang, and T. S. Huang, "Maximum margin GMM learning for

facial expression recognition," *2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013*, pp. 1–6, Apr. 2013.

[135] H. Tang, M. Hasegawa-Johnson, and T. Huang, "Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors," in *2010 IEEE International Conference on Multimedia and Expo*, 2010, pp. 1202–1207.

[136] S. Moore and R. Bowden, "Multi-view pose and facial expression recognition," *Br. Mach. Vis. Conf. BMVC 2010 - Proc.*, pp. 1–11, 2010.

[137] C. Shan, S. Gong, and P. P. McOwan, "Dynamic Facial Expression Recognition Using A Bayesian Temporal Manifold Model," *Procdings Br. Mach. Vis. Conf. 2006*, pp. 31.1–31.10, 2006.

[138] H. Soyel and H. Demirel, "Improved SIFT matching for pose robust facial expression recognition," in *Face and Gesture 2011*, 2011, pp. 585–590.

[139] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, and T. S. T. Huang, "Multi-view facial expression recognition," *Autom. Face Gesture Recognition, 2008. FG'08. 8th IEEE Int. Conf.*, pp. 1–6, Sep. 2008.

[140] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction.," *2003 Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 5, pp. 53–53, Jun. 2003.

[141] S. Berretti, A. Del Bimbo, P. Pala, B. Ben Amor, D. Mohamed, B. Ben Amor, and M. Daoudi, "A set of selected SIFT features for 3D facial expression recognition," *Proc. - Int. Conf. Pattern Recognit.*, pp. 4125–4128, 2010.

[142] O. Rudovic, I. Patras, and M. Pantic, "Facial expression invariant head pose normalization using gaussian process regression," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 28–33, 2010.

[143] J.-G. Wang, J. Li, C. Y. Lee, and W.-Y. Yau, "Dense SIFT and Gabor

descriptors-based face representation with applications to gender recognition," in *2010 11th International Conference on Control Automation Robotics & Vision*, 2010, pp. 1860–1864.

[144] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "Emotion Recognition from Arbitrary View Facial Images," 2010, pp. 490–503.

[145] O. Tuzel, F. Porikli, and P. Meer, "Region Covariance: A Fast Descriptor for Detection and Classification," 2006, pp. 589–600.

[146] U. Tariq, J. Yang, and T. S. Huang, "Multi-view Facial Expression Recognition Analysis with Generic Sparse Coding Feature," 2012, pp. 578–588.

[147] Jianchao Yang, Kai Yu, Yihong Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1794–1801.

[148] Y. W. Chen and C. J. Lin. "Combining SVMs with various feature selection strategies." *Feature extraction*. Springer Berlin Heidelberg, 2006. 315-324.

[149] M. K, V. Govindarajan, S. K. V V S, and R. S, "Face Recognition using Block-Based DCT Feature Extraction," *J. Adv. Comput. Sci. Technol.*, vol. 1, no. 4, pp. 266–283, Aug. 2012.

[150] Y. Tong, R. Chen, and Y. Cheng, "Facial expression recognition algorithm using LGC based on horizontal and diagonal prior principle," *Optik (Stuttg).*, vol. 125, no. 16, pp. 4186–4189, 2014.

[151] M. Pietikäinen, "Emotion recognition from facial images with arbitrary views," *Bmvc2013*, vol. 1, no. 2, pp. 2, 2013.

[152] S. Moore and R. Bowden, "The effects of Pose on Facial Expression Recognition," *Procdings Br. Mach. Vis. Conf. 2009*, pp. 79.1–79.11, 2009.

[153] H. Soyel, U. Tekguc, and H. Demirel, "Application of NSGA-II to feature selection for facial expression recognition," *Comput. Electr. Eng.*, vol. 37, no.

6, pp. 1232–1240, 2011.

[154] Xiaoyang Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[155] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[156] V. Takala, T. Ahonen, and M. Pietikäinen, "Block-based methods for image retrieval using local binary patterns," *Scand. Conf. Image Anal.*, vol. 3540, pp. 882–891, 2005.

[157] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[158] Y. Hu, Z. Zeng, Lijun Yin, Xiaozhou Wei, Jilin Tu, and T. S. Huang, "A study of non-frontal-view facial expressions recognition," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[159] N. Hesse, T. Gehrig, H. Gao, and H. K. Ekenel, "Multi-view Facial Expression Recognition using Local Appearance Features," *Int. Conf. Pattern Recognit.*, no. Icpr, pp. 3533–3536, 2012.

[160] X. Huang, G. Zhao, M. Pietikäinen, and W. Zheng, "Expression Recognition in Videos Using a Weighted Component-Based Feature Descriptor," in *Image Analysis*, Springer Berlin Heidelberg, 2011, pp. 569–578.

[161] J.-J. Wong and S.-Y. Cho, "A face emotion tree structure representation with probabilistic recursive neural network modeling," *Neural Comput. Appl.*, vol. 19, no. 1, pp. 33–54, Feb. 2010.

[162] A. Ramirez Rivera, J. Rojas Castillo, and O. Oksam Chae, "Local Directional Number Pattern for Face Analysis: Face and Expression Recognition," *IEEE*

Trans. Image Process., vol. 22, no. 5, pp. 1740–1752, May 2013.

[163] W. Gu, C. Xiang, Y. V Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognit.*, vol. 45, pp. 80–91, 2011.

[164] S. H. Lee, K. N. K. Plataniotis, and Y. M. Ro, "Intra-Class Variation Reduction Using Training Expression Images for Sparse Representation Based Facial Expression Recognition," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 340–351, Jul. 2014.

[165] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Syst. Man. Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[166] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Comput. Vis. Image Underst.*, vol. 115, no. 4, pp. 541–558, 2011.

[167] F. Dornaika, A. Moujahid, and B. Raducanu, "Facial expression recognition using tracked facial actions: Classifier performance analysis," *Eng. Appl. Artif. Intell.*, vol. 26, no. 1, pp. 467–477, 2013.

[168] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas, "Facial expression recognition using encoded dynamic features," *Comput. Vis. Pattern Recognit.*, pp. 1–8, 2008.

[169] Y. Sun and L. Yin, "Facial Expression Recognition Based on 3D Dynamic Range Model Sequences," *Eccv*, pp. 58–71, 2008.

[170] A. Ryan, J. Cohn, and S. Lucey, "Automated Facial Expression Recognition System," *43rd Annu. 2009 Int. Carnahan Conf. Secur. Technol.*, pp. 172–177, 2009.

[171] M. Pantic and I. Patras, "Detecting facial actions and their temporal segments in nearly frontal-view face image sequences," *2005 IEEE Int. Conf. Syst. Man*

Cybern., vol. 4, pp. 3358–3363, 2005.

[172] N. L. Etcoff and J. J. Magee, "Categorical perception of facial expressions," *Cognition*, vol. 44, no. 3, pp. 227–240, 1992.

[173] P. E. Griffiths and E. Walsh, *Emotion and Expression*, Second Edi., vol. 7. Elsevier, 2015.

[174] K. Scherke, "Emotion: History of the Concept," *Int. Encycl. Soc. Behav. Sci.*, vol. 10, pp. 139–143, 2015.

[175] V. Shuman and K. Scherer, "Psychological Structure of Emotions," *Int. Encycl. Soc. Behav. Sci. 2nd Ed.*, vol. 7, pp. 526–533, 2015.

[176] E. Hudlicka, "To feel or not to feel: The role of affect in human–computer interaction," *Int. J. Hum. Comput. Stud.*, vol. 59, no. 1, pp. 1–32, 2003.

[177] V. Bettadapura, "Face," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, 2000.

[178] E. G. Blanchard, et al. "Affective artificial intelligence in education: From detection to adaptation." *AIED*. Vol. 2009. 2009.

[179] P. Ekman, "Expression and the nature of emotion," in *Approaches to emtoion*, Erlbaum, 1984, pp. 319–343.

[180] C. E. Izard, *Human emotions*. New York: Plenum, 1977.

[181] R. S. Woodworth and H. Schlosberg, *Experimental psychology*. Oxford and IBH Publishing, 1954.

[182] H. Schlosberg, "Three dimensions of emotion.," *Psychol. Rev.*, vol. 61, no. 2, p. 81, 1954.

[183] M. B. Arnold and J. A. Gasson, "Feelings and emotions as dynamic factors in personality integration," *Hum. Pers.*, pp. 294–313, 1954.

[184] K. R. Scherer, and H. Ellgring. "Multimodal expression of emotion: Affect programs or componential appraisal patterns?" *Emotion,* Vol. 7.1, pp. 158-171, 2007.

[185] K. R. Scherer, "Appraisal theory.," *Handbook of cognition and emotion*. pp. 637–663, 1999.

[186] C. A. Smith, "Dimensions of appraisal and physiological response in emotion," *Journal of personality and social psychology,* Vol. 56.3, pp. 339-353, 1989.

[187] K. R. Ellsworth, P. C., & Scherer, P. Ellsworth, and K. Scherer, "Appraisal processes in emotion," *Handbook of affective sciences*, 2003, pp. 572–595.

[188] W. S. Torgerson, "Theory and methods of scaling," Oxford, England, Wiley, 1958, pp. 460.

[189] J. Frois-Wittman, "The judgment of facial expression.," *J. Exp. Psychol.*, vol. 13, no. 2, p. 113, 1930.

[190] H. Scholsberg, "A scale for the judgment of facial expressions.," *J. Exp. Psychol.*, vol. 29, no. 6, p. 497, 1941.

[191] R. P. Abelson and V. Sermat, "Multidimensional scaling of facial expressions.," *J. Exp. Psychol.*, vol. 63, no. 6, p. 546, 1962.

[192] L. Feldman Barrett and J. A. Russell, "Independence and Bipolarity in the Structure of Current Affect," *J. Pers. Soc. Psychol.*, vol. 74, no. 4, pp. 967–984, 1998.

[193] M. S. M. Yik, J. A. Russell, and L. F. Barrett, "Integrating four structures of current mood into a circumplex: Integration and beyond," *J. Pers. Soc. Psychol.*, vol. 77, pp. 600–619, 1999.

[194] D. Watson, D. Wiese, J. Vaidya, and A. Tellegen, "The two general activation systems of affect: Structural findings, evolutionary considerations, and

psychobiological evidence.," *J. Pers. Soc. Psychol.*, vol. 76, no. 5, pp. 820–838, 1999.

[195] R. J. Larsen and E. Diener, "Promises and problems with the circumplex model of emotion," *Emotion*, Thousand Oaks, CA, US: Sage Publications Inc., 1992, pp. 25-29.

[196] R. E. Thayer, *The biopsychology of mood and arousal*, Oxford University Press, 1989.