Facial expression recognition in the spatio-temporal domain

Xiang, Tuo Wen

2009

Xiang, T. W. (2009). Facial expression recognition in the spatio-temporal domain. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/47476>

<https://doi.org/10.32657/10356/47476>

Nanyang Technological University

Downloaded on 26 Apr 2023 23:22:37 SGT

FACIAL EXPRESSION RECOGNITION IN THE SPATIO- TEMPORAL DOMAIN

XIANG TUOWEN

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2009

Abstract

In recent years, automatic Facial Expression Recognition (FER) has been extensively studied and has gained much research interest due to its many useful applications. However, to the best of our knowledge, the proposed algorithms all suffer from some limitations, which make the systems work inaccurately and unstably in the real scenarios. In this thesis, we aim to propose and investigate a full automatic system to give effective and accurate classification of facial expressions.

A sequence of expression images instead of just one image is analyzed in the proposed system. Each sequence displays one expression, starting from a neutral state with no expression and ending at a maximum muscle distortion state. We proposed a feature descriptor for facial component localization, after which, the faces in the image sequence can be easily normalized into the same pose and scale. For facial expression recognition system, we favor descriptors that are based on local information rather than global expression pattern since local descriptors can potentially be applicable to occluding and cluttering environment. The proposed employs a novel local feature descriptor of Fourier Transform components to tolerate lighting changes and extract sub-space information from the 3D Spatio-Temporal space. Three kinds of local information descriptors are proposed in this thesis: the

Wave Shape Spatio-Temporal (*WST*) descriptor, 1D Fuzzy Spatio-Temporal (*1D-FST*) descriptor and 2D Fuzzy Spatio-Temporal (*2D-FST*) descriptor. The transformed frequency information is further processed by different classifiers to generate a Spatio-Temporal model for each expression type. Unknown input expressions will be matched to the models for classification. Modified Hausdorff Distance (MHD) is employed to compute the dissimilarity scores and the minimal score indicates the corresponding expressions of the input data.

The proposed technique has been tested with the Cohn-Kanade and MMI database. The experiments were carried out in subject independent way with leave one subject out cross validation and cross database validation. In the experiments, we have also explored the relationship between the system performance and the parameters settings. We get some interesting observations.

In this work, the development of the research of facial expression recognition has been reviewed. Some limitations of the current systems have been investigated and analyzed. An in-depth study of facial expression recognition on the Spatio-Temporal domain has been carried out and a novel, automatic and effective system has been built to overcome some of the existing limitations. A series of thorough experiments have been conducted and the results are promising and encouraging. At the end, the direction of future work has also been suggested by analyzing the error cases and the drawback of the proposed system.

Acknowledgements

First and foremost, I would like to express my most gratitude to my supervisor, Dr. Maylor K. H. Leung for the time and effort in supervising this thesis. His personal enthusiasm, honest criticism, professional guidance and encouragement help me a lot throughout my PhD study. Without his patience and well-meaning criticism, the thesis could not be finished.

Thank my friends for helping in the collection of experiment data. They are Guo Yufeng, Zhang Bingru, Deng Shusheng, Han Yi, Xie Jun, Guo Dagang, Jia Hui.

My thanks also go to the technicians of Database Technology Lab, Mr. Teo Choo Eng and Ms. Eng Hui Fang for all their kind and generous assistance.

I would also like to thank Nanyang Technological University (NTU) for providing the research scholarship during his candidature in this research program.

Thank Mr. CherKeng Heng for his kind help and suggestion in my thesis revision.

My experiments data comes from the Cohn-Kanade database and MMI database. Thank the owners to offer such good face data.

Thank my examiners for their detailed and valuable comments.

Last and most, my sincere thanks and gratitude are given to my parents for their constant support, encouragement and blessings.

Contents

Chapter 1	Introduction.....	1
1.1	Background.....	1
1.2	Motivation.....	4
1.3	Proposed Approach.....	6
1.4	Objectives.....	7
1.5	Testing Data.....	10
1.6	Thesis Organization.....	11
Chapter 2	Literature Review.....	13
2.1	Face Detection.....	14
2.1.1	Feature Level.....	15
2.1.2	Weak Classifier Level.....	16
2.1.3	Strong Classifier Level.....	17
2.1.4	System Structure Level.....	18
2.2	Facial Components Localization.....	18
2.2.1	The Selected Features Representation.....	21
2.2.2	Machine Learning Methods.....	22
2.3	Facial Expression Recognition.....	23
2.3.1	Facial Expression Data Extraction.....	24
2.3.2	Facial Expression Classification.....	27
2.4	Implementation Tools.....	29
2.4.1	Fourier Transform.....	29
2.4.2	Fuzzy C Means.....	30
2.4.3	Hausdorff Distance.....	32
2.5	Summary.....	33
Chapter 3	Pre-processing.....	34
3.1	Input Data Image.....	35
3.1.1	Cohn-Kanade Database.....	35
3.1.2	MMI Face Database.....	36
3.2	Pre-processing.....	37
3.2.1	Face Detection.....	37

3.2.2	Facial Component Localization	38
3.2.2.1	System Overview	39
3.2.2.2	Training Samples.....	40
3.2.2.3	Feature Extraction	41
3.2.2.4	Classifiers Training Procedures.....	43
3.2.2.5	Facial Components Localization Procedures	46
3.2.2.6	Facial Component Localization on Occluded Face	49
3.2.2.7	Discussion	51
3.3	Facial Area Extraction & Normalization.....	52
3.4	Frame Selection.....	53
3.5	Summary	54
Chapter 4	Facial Expression Recognition.....	56
4.1	Feature Descriptor Representation	57
4.1.1	1D Fourier Transform Extraction	57
4.1.2	2D Fourier Transform Extraction	60
4.2	Classifiers Processing.....	63
4.2.1	Wave Shape Approach	66
4.2.1.1	Feature Extraction	66
4.2.1.2	Model Construction & Recognition	68
4.2.2	Fuzzy C Means Approach	70
4.2.2.1	Feature Extraction	70
4.2.2.2	Expression Model Construction	73
4.2.2.3	Facial Expression Recognition.....	74
4.3	Summary	78
Chapter 5	Experimental Results	79
5.1	Facial Component Localization	80
5.2	Facial Expression Recognition.....	83
5.2.1	1D_FC_WaveShape System Results.....	84
5.2.1.1	Expression Recognition.....	84
5.2.1.2	Comparison	85
5.2.2	1D_FC_FCM System Results	86
5.2.2.1	Expression Recognition.....	86
5.2.2.2	Empirical Optimal Setting.....	88
5.2.2.3	The Effects of Resolution and Number of Frames	90
5.2.3	2D_FC_FCM System Results	93
5.2.3.1	Expression Recognition.....	93
5.2.3.2	Empirical Optimal Setting.....	94
5.2.3.3	The Effects of Number of Frames	97
5.2.4	Cross Database Validation.....	100
5.2.5	Bidirectional Test	101
5.2.6	Comparison with Other Results	103
5.3	Summary	106

Chapter 6 Conclusion 107

6.1 Work Summary 107

6.2 Contributions..... 109

6.3 Future Work..... 111

6.3.1 2D Face Alignment 111

6.3.2 3D Face Alignment 112

6.3.3 Facial Expression Recognition..... 113

6.4 Thesis Conclusion 114

References..... 116

Publication List..... 126

Appendix I: Examples of Facial Components Localization..... 127

Successful Localization..... 127

Unsuccessful Facail Components Localization..... 134

List of Figures

Fig. 1-1 Different expression examples from the Cohn-Kanade database.	11
Fig. 1-2 Thesis organization.	12
Fig. 2-1 Relationship between time and frequency domain.	30
Fig. 2-2 Examples of Non Fuzzy set and Fuzzy set.	31
Fig. 3-1 The system pre-processing flowchart.	34
Fig. 3-2 An example from the Cohn-Kanade database.	36
Fig. 3-3 The 9 facial components for normalization.	40
Fig. 3-4 Some examples of eye inner corners positive training samples.	41
Fig. 3-5 Left: one point set p_k on the face; Right: Gabor filter bank $G \{g_0 \dots g_{11}\}$	42
Fig. 3-6 Six point sets for Adaboost selection $\{p_0 \dots p_5\}$	42
Fig. 3-7 Examples of left/right inner eye corner projected scores.	44
Fig. 3-8 The equation for parameters (B) estimation.	45
Fig. 3-9 Facial component localization system.	47
Fig. 3-10 An example of pyramid images.	47
Fig. 3-11 An example of facial area normalization.	53
Fig. 4-1 Structure of the FER system.	57
Fig. 4-2 An expression sequence seen in 3D space.	58
Fig. 4-3 An example of input data of 1D Fourier Transform.	59
Fig. 4-4 An example of output data of 1D Fourier Transform.	60
Fig. 4-5 Examples of $f'(x,y,t)$ in 3D space.	62
Fig. 4-6 Examples of the expression Fourier spectra.	64
Fig. 4-7 Examples of different types of mouth movements with expressions	65
Fig. 4-8 Examples of $f(x, y, t)$ and Fourier Components.	69
Fig. 4-9 Typical examples of the convergence of u_{ijk}	75
Fig. 4-10 Typical examples of the convergence of μ_k	75
Fig. 4-11 Membership matrices of the six basic expressions.	76
Fig. 5-1 Different degrees of mouth movements for the 'Smile' expression.	88
Fig. 5-2 Recognition rates with respect to c and N%.	89
Fig. 5-3 Recognition rate with respect to different resolutions.	91
Fig. 5-4 Recognition rates with respect to different numbers of frames.	92
Fig. 5-5 Examples of input images with different resolutions.	92
Fig. 5-6 The recognition rates of 1D and 2D Fourier Transforms with respect to the numbers of frames.	99

List of Tables

Table 3-1 Eigenvalues derived from a set of face images.....50

Table 4-1 Expression types with respect to *i*.....74

Table 5-1 Numbers of facial expression sequences of the Cohn-Kanade database.....79

Table 5-2 Experimental results for facial components localization.81

Table 5-3 Experimental results for Cohn-Kanade database.85

Table 5-4 Subject dependent and independent system performance.....85

Table 5-5 The best recognition results of 1D FST.....88

Table 5-6 Summary of the system parameters.88

Table 5-7 Recognition rates with respect to *c* and *N*%.89

Table 5-8 Parameter setting for different resolution.....91

Table 5-9 Parameter Setting for different number of frames.91

Table 5-10 Summary of the system parameters.94

Table 5-11 The best recognition results of 2D FST.....94

Table 5-12 Recognition rates with respect to *c* and *N*% when *N_p*=10.....95

Table 5-13 Recognition rates with respect to *c* and *N*% when *N_p*=20.....95

Table 5-14 Recognition rates with respect to *c* and *N*% when *N_p*=30.....95

Table 5-15 Recognition rates with respect to *c* and *N*% when *N_p*=40.....96

Table 5-16 Recognition rates with respect to *c* and *N*% when *N_p*=50.....96

Table 5-17 Recognition rates with respect to *c* and *N*% when *N_p*=100.....96

Table 5-18 The optimal setting of parameters with respect to the numbers of frames.....99

Table 5-19 Cross database validation results.100

Table 5-20 Bidirectional testing results.....102

Table 5-21 Results of recent work using static image.....104

Table 5-22 Comparisons of recent work using sequences of images.....105

Chapter 1 Introduction

Automatic facial expression recognition has attracted significant interest in the scientific community due to its many useful applications [1][2][3][4][5]. A fully automatic and effective approach is proposed in this thesis to analyze a sequence of expression images instead of just one image. The proposed employs a novel local descriptor of Fourier Transform components to tolerate lighting changes and extract sub-space information from the 3D Spatio-Temporal space. The transformed frequency information is further processed by different classifiers to generate a Spatio-Temporal model for each expression type. Unknown input expressions will be matched to the models for classification.

1.1 Background

Facial expressions are intrinsic to communication and play a significant role in our social and emotional lives [6][7][8][123]. It is one of the most important, naturally preeminent means for human beings to display emotions, to stress and clarify the verbal information, to signal comprehension, disagreement, and intentions, and to regulate interactions with the environment and other people in the vicinity. In the

interaction between humans, adequate feedback like speech, facial expression, and body gestures are essential modalities since these components satisfy certain communication expectations [9]. Furthermore, the human face constitutes a source of informative social signs that allow good communication expectation of response. Indeed, it has been stated by Mehrabian [10] that the textual content of a spoken message contributes only for 7% of the message effect while facial expressions and voice intonation contribute for 55% and 38% of the global message impact respectively. Psychological studies have also indicated that humans interact with each other mainly through speech, but also through display of emotions for emphasis purpose [11]. One of the important way humans display emotions is through facial expressions. The communicative power of expressions makes it a very active research topic and has attracted a significant interest in the scientific community. In [12], the facial expressions are concluded to provide useful information in different areas such as:

- Business: Facial expressions are important in business negotiations and personal decisions. It can indicate clues of the concealed emotion when the plan and action are provided in words.
- Education: The teacher's facial expressions influence the effect of the pupils learning and the pupils' facial expressions can reflect and inform the teacher of the need to adjust the instructional message.
- Medicine: Facial expressions can be useful in studies of the psychological state of the patient. It can provide not only the diagnostic information relevant

to depression, mania, schizophrenia, and other less severe disorders, but also the information relevant to monitoring response to treatment. It is also essential to evaluate psychological state of the preverbal infants and mute disabled person.

- Criminal Identification: Facial expressions play a crucial role in establishing or detracting the credibility of the suspect.
- Human-computer Interactions: Facial expressions can provide a way to communicate basic information about needs and demands of the users to computers.

With the rapid development of computer technology, there is a need to make automatic systems in above areas to ease and simplify the human operation. Based on this observation, automatic facial expression recognition (FER) [13][14][15][16][17][18] becomes a crucial pre-process and is required for these systems to work more accurately and provide better performance. Moreover, the effective solution of FER would have the following useful application in computer science, such as interactive video, indexing and retrieval of image and video databases, lip reading, speech analysis, image understanding, video conference, face and visual speech animation, virtual character animation, and affective computing [19]. Among them, the most important potential application is human-computer interaction. It is crucial for computers to be able to interact with the users, in a way similar to human-to-human interaction [20]. More accurate feedback can be given by the computer or robot if they can identify the expressions of the users successfully.

Recently there has been a significant amount of research on FER in computer vision and graphics [21][22][23][24][25][26][27][28]. Psychological studies indicate that there are six facial expressions associated universally with six distinct emotions as smile, sadness, surprise, fear, anger and disgust [29][30]. Many FER systems are examined using these six basic expressions [31][120]. Thus, we also adopt this criterion to examine our system in this thesis.

1.2 Motivation

Facial expression recognition has many interesting potential applications. However, due to different age, gender and external objects such as glasses and hair, automatic facial expression recognition is still a complex challenge. Also, faces appear different because of pose and lighting condition changes. Variations such as these should be considered and handled by an automatic FER system [32]. To solve this kind of problems, current approaches assume some precondition to be satisfied before the identification of facial expression. Robust face detection and tracking are of great importance for some FER systems. Moreover, approximate face position can not satisfy the requirement of further face analysis. Even after successful face detection, interesting facial components that are usually utilized for facial expression analysis, such as the mouth corners, eyebrows, eyes, nostrils, chin, etc., need to be located [33].

While face detection achieves huge success by Viola's work [34], currently facial

feature localization still meets some problems to get both high accuracy and fast speed at the same time. Viola has proven that the Adaboost is a powerful tool for solving practical pattern recognition problems based on learning from data in [34]. Some researchers try to use Viola's system (Haar Feature) for facial feature localization [35]. The result is not good enough to reach the requirement. It may be because that the features they used are not discriminable enough [36]. We observed that Gabor features are ubiquitously used in computer vision area and achieve good results. Motivated by this, we adopted Viola's system to find the rough position of face and proposed a new feature representation based on Gabor filters for facial component positions localization.

For facial expression recognition, as mentioned previously, the psychological study have shown that temporal information can reveals more relevant information about the underlying emotional states than a single static image. It is essential to extract some expression feature descriptor from the muscle movement to represent the discrimination between different expressions. In this thesis, we are interested to build a robust, accurate and effective system which can tolerate some variations of environment changes and classify the expressions quickly and correctly.

In the following part, Chapter 1.3 describes our system briefly and Chapter 1.4 illustrates some of our concerns and proposed solution.

1.3 Proposed Approach

In this thesis, we proposed a novel Spatio-Temporal feature descriptor for real time facial expression recognition. The proposed system first detects the face and facial component positions. Based on these positions, the face can be automatically normalized into the same scale. After that, the system employs Fourier Transform to change a facial expression sequence from the Spatio-Temporal domain into the Spatio-Frequency domain. This is followed by a description process that extracts and describes the local information. Six expression models can then be constructed. An input expression will be matched to these models using Hausdorff Distance [37][38] for classification. Two kinds of local information classifiers are used in this thesis: the Wave Shape (WS) classifier and Fuzzy C means (FCM) classifier [39][40]. WS classifies an input expression according to the shape of intensity changes of every pixel after Fourier Transform. FCM is different from WS. It gives fuzzy membership to each input expression pattern, which can determine the intrinsic division in a set of unlabeled data and find representatives for homogeneous groups. Unlike hard c-means algorithm which executes a sharp classification in which each object is either assigned to a class or not, FCM can make an object belong to several classes at the same time but with different degrees between 0 and 1. Since some expressions might look alike at certain facial regions, it is hard to put clear boundaries between different expressions. Thus, FCM is used since it can give intermediate memberships instead of hard boundaries.

1.4 Objectives

The goal of the work presented here is to develop a recognition system that can tackle or tolerate the current FER system limitations to certain extent with the following focuses.

1. Automatic face normalization: Input faces with different sizes and poses can create problem for current FER system. When the face could be located exactly and normalized into the same scale and pose, the recognition step afterwards would not be so complicated [41]. To solve this problem, some existing systems normalize the face data manually. However, this would involve human interaction and make the system not fully automatic. Here we propose a facial feature localization classifier trained by Adaboost. This classifier scans the face image with different scales and locates 9 facial component positions accurately. Based on the distance between these 9 facial components, the input face can be automatically normalized for further processing.
2. Light condition variation: Face image may appear different for the same person with the same pose when the environment illumination condition changes. We propose to use face Spatio-Temporal information to describe the expression instead of static face image, since we believe that each expression have its own motion pattern. It is also based on the psychological study [42] that temporal information can reveals more relevant information about the underlying emotional states than a single static image. For example, it is

difficult to distinguish a static image of a slightly smiling face from a face with neutral expression. Compared with static image feature, the temporal motion pattern is less influenced by the light variation. In our system, Fourier Components (1D or 2D) are used to form the expression feature descriptor. Since Fourier Components record the frequency information of the Spatio-Temporal pattern, the system based on Fourier Components is more robust to the light condition changes.

3. Local feature descriptor: We favor descriptors that are based on local information rather than global expression pattern since local descriptors can potentially be applicable to occluded and cluttered environment. The search needs to tackle the following issues:

- a) *Dimension*: A sequence of images spans a 3-dimensional space of position (x - y) and time (t). Hence, the descriptor can be built on:

- 1-dimensional data: intensity changes with respect to t at each pixel.
- 2-dimensional data: intensity changes with respect to t at each vertical or horizontal line.

Appropriate consideration should be given to find the suitable combination. Both the 1-dimensional and 2-dimensional descriptors are proposed in this thesis.

- b) *Size*: The size of a descriptor is important since it specifies the amount of local information and affects the training process, i.e. larger size gives rise to more combinations of patterns that can be described by a descriptor and

hence more training examples are needed. We have experimented with different sizes in order to tune and understand this parameter. The rationale can be extended to other recognition systems that employ training.

c) *Effectiveness / efficiency*: It is likely that more information can lead to better result if the information is relevant. On the other hand, if redundant or irrelevant information is included, the result can deteriorate and take much longer time to compute. Hence, it is of interest here to experiment with the descriptor size and other parameters to find and suggest the right amount of information to obtain optimal or near optimal result that can be computed efficiently.

4. Fuzzy Processing: Human changing their expression is a natural, continuous and smooth process. From non-expression to a maximum expression, it is very hard to say from a certain frame, the person starts to play one expression. Even when people change one expression to another, the boundary between is blur and illegible. Motivated by this, we introduce Fuzzy classifier into FER field. Fuzzy C Means is used since it can give intermediate memberships for each feature descriptor instead of hard boundaries.
5. Misalignment toleration: In general, it is hard for face images to be 100% normalized or aligned into the same scale. Resizing or rotating operation always brings some small misalignment to the image. Hausdorff distance (HD)

is employed to tackle misalignment. HD searches in a small neighborhood and find the most similar patch.

1.5 Testing Data

The proposed systems are tested by the popularly used Cohn-Kanade facial expression database [43] and MMI face database [114]. Cohn-Kanade facial expression database consists of 96 subjects, 391 expression sequences with resolution of 640×480 and frame rate at 30 frames per second. Many other systems had evaluated their performances on this database. This facilitates comparison of our system to other expression recognition systems. It captures full-face frontal views under constant illumination at relatively the same scale of facial area. Each expression recording starts at a neutral expression with no expression and ends at the peak of the muscle distortion. Each subject plays one or more than one of the basic six expressions, but not necessary all the six expressions. Some examples are displayed in Fig. 1-1.

The MMI Face Database is an easily accessible, comprehensive, benchmark database of facial expression [114]. It contains more than 1500 samples of both static images and image sequences of faces both in frontal and in profile view, displaying various facial expressions of emotion, single AU activation, and multiple AU activation. In this thesis, we use MMI face data for the cross database validation and bidirectional evaluation.

Fig. 1-1 Different expression examples from the Cohn-Kanade database.

1.6 Thesis Organization

The rest of the thesis is organized as shown in Fig. 1-2. Chapter 2 gives a detail review of facial expression recognition methods. Chapter 3 presents the system pre-processing. Following this, the proposed approaches (including the feature representation, classifiers building and expression recognition) are described in chapters 4 with details. Chapter 5 shows the experimental results and presents the comparison with other systems. The conclusion and future work are described in chapter 6.

Fig. 1-2 Thesis organization.

Chapter 2 Literature Review

In recently years, automatic Facial Expression Recognition (FER) has been extensively studied and has gained much research interest [44][45][46][47][48]. The reasons for this interest are multiple, but mainly due to developments accomplished in related research areas such as face tracking, face detection and face recognition as well as the recent availability of the equipments with relatively cheaper price and more powerful computation ability. Another important reason fostering this research interest is that a lot of potential applications of FER can be envisaged in the near future, including interactive video, indexing and retrieval of image and video databases, image understanding, lip reading, speech analysis, video conference, face and visual speech animation, virtual character animation, and affective computing. Thus, a wide range of algorithms has been applied to the automatic FER problem [49][50][51]. However, although some systems have achieved promising results, due to different age, gender and external objects such as glasses and hair, automatic FER is a complex challenge. Furthermore, faces appear different because of pose and lighting condition changes. Variations such as these should be considered and handled by an automatic FER system. Among the proposed systems, many of them have an assumption that the faces in an image or an image sequence have been localized and

normalized. To build a fully automated system that can analyze the information contained in face images, robust and efficient face detection and facial features localization algorithms are required. Hence in this chapter, some face related systems are reviewed. Furthermore, Fourier Transform [52][53][54][55], Hausdorff Distance [56][57][58] and Fuzzy C Means [59][60][61][62][63] are employed for classification of facial expression in this thesis. A brief description about these methods is also presented here.

2.1 Face Detection

After a decade of efforts, face detection research has achieved great advances. It is assumed that human faces are highly correlated. Even some of variation and complexity brought by the changes in facial appearance, lighting and expression, it is believed that face can be described by some low-dimensional features derived from a set of prototype face images [64].

Numerous methods have been proposed for face detection. M. Yang et. al. gave a detailed survey [65] about the face detection methods in 2002. However, the main break through is due to Viola's Boosted Cascade Framework [34], which is clearly distinguished from previous approaches in its ability to detect human faces extremely accurately and rapidly. It is believed that this remarkable performance owes to the fast speed of Haar-like feature calculation based on the integral image, the high accuracy of boosted classifiers, and the asymmetric cascade structure for decision making [66].

Viola's system solved the two key problems for face detection: the detection rate and the processing speed and achieved great success. However, the framework meets some problems when dealing with the multi-view faces detection, whose patterns are more diverse and asymmetric [66]. It is found that the compact and symmetry Haar-like features are not as adaptable as they were with frontal faces. From then on, more and more derivatives of Viola's work have been proposed to improve the face detection system's accuracy and robustness. The main modifications are concentrated on 4 levels: feature level, weak classifier level, strong classifier level and system structure level. In the following part, the derivate works are review according to this classification.

2.1.1 Feature Level

In [67], Rainer and Jochen extended Viola's rapid object detection framework in two ways. Firstly, their basic and over-complete set of Haar-like feature was extended by an efficient set of 45° rotated features, which added additional domain-knowledge to the learning framework and which was otherwise hard to learn. These novel features can also be computed rapidly at all scales in constant time. Secondly, they derived a new post-optimization procedure for a given boosted classifier that improves their system performance. Compared with Viola's feature set, rotated Haar-like features were added and the author claimed that this increased the capacity of the learning algorithm.

In [68], a new kind of features has been proposed and been integrated in the object detection scheme. Unlike Viola's work which compared the sums of pixel values in rectangular regions, the proposed feature was based on statistical examination of individual pixel values. The advantage is that the feature is insensitive to variance normalization, since the answer "yes" or "no" only depends on the relations between pixels values and not on some quantitative thresholds.

2.1.2 Weak Classifier Level

In [69], Wu Bo et. al. brought the ideas of real Adaboost and LUT (Look Up Table) into Viola's face detection system and achieved significant improved performance compared with previous results. Instead of using the Boolean weak classifiers, Wu's system output a real value. By dividing feature space into many sub-regions with equal width, they developed a piece-wise function (LUT) to approximate complex distribution of training samples according to the real Adaboost output.

Viola reported that features which were selected in later rounds yielded error rates between 0.4 and 0.5 while features selected in early rounds had error rates between 0.1 and 0.3 [34]. By observing this, Mita et. al. claimed that "too weak" classifiers did not contribute to improving the generalization performance [70]. To solve this problem, it is necessary to find some more distinctive features, which can capture the structural similarities among the face classes. Therefore, the authors proposed a new

kind of features, called joint Haar-like features, for detecting faces in images. In contrast with the method proposed by Viola, Mita's system used multiple Haar-like features in each weak classifier. The authors also claimed that Gabor features were strong candidates for further improvement of the classification performance.

2.1.3 Strong Classifier Level

Viola used Adaboost to combine several "weak" classifiers to compose a "strong" classifier. In [71], Liu and Shum observed that Adaboost left two questions unsolved. First, how to best combine the weak classifiers in terms of the coefficients; second, how to choose the best weak classifiers or features. Liu then proposed a classification framework called KL Boosting. It solved the first problem by iteratively tuning the coefficients to minimize the recognition error. It ensured the recognition error would not increase as more features involved. To solve the second problem, KL Boosting found KL features that maximize the symmetric KL divergence between two classes, which corresponds to the projected margin. In other words, KL Boosting used optimal features to constitute an optimal classifier.

In [72], a novel framework for rapid object detection was presented. In this framework, boosting cascade and bootstrap training were integrated into a single learning procedure ("chain structure"), which not only provided a theoretical foundation for cascade training, but also improved classifier performance by incorporating historical knowledge of cascade learning. The authors claimed that the

resulting classifier consists of fewer weak classifiers yet achieves lower error rates than boosting cascade in both training and testing.

2.1.4 System Structure Level

In [73], a tree-structure system for face detection was presented, which adopted the coarse-to-fine strategy to divide the entire face space into smaller and smaller subspaces. Four such WFS (Width-First-Search) tree-structured detectors worked as a whole one. An experimental result was reported that a rotation invariant face detection system was implemented by using this tree-structure and achieved real time performance with high accuracy.

2.2 Facial Components Localization

Facial components localization, as an important research area in human face analyzing, is specialized in searching for facial components (i.e. eyes, nose, mouth, and eyebrow) within a given region in an image or image sequence. The applications can be found in many areas, such as face detection, face recognition, facial expression recognition and reconstruction of face cartoon, etc. However, facial component localization is complicated and also faces the similar problems that are encountered with face detection, such as lighting, pose, expression variations, etc. The moustache can be easily confused to be a mouth lip, while glasses sometimes bring much trouble

for human eyes detection. Even having found the gross positions of faces, robust and precise facial feature locating is still a challenging task.

Among all the facial features, eyes are most frequently detected for face alignment. The eye center positions are detected to accurately locate and normalize the face before further processing. Eye detection becomes more and more important since it is obvious to see that their position accuracy may influence the overall system performance.

In [36], the authors discussed the relationship between face recognition system performance and eye detection accuracy. To observe how the recognition rate varies according to eye localization error, the eye positions of the ground truth were artificially changed with random noise. Face recognition was then performed based on the inaccurate eye positions. The impact of eye detection on recognition accuracy was illustrated and it was found that eye localization errors significantly affect the recognition accuracy.

Other researchers have also drawn similar conclusions. In [74], the face recognition algorithms with automatic and manual alignment were both evaluated. The face recognition algorithms with manual eye coordinates were referred as “partially automatic algorithms” while “fully automatic algorithms” as automatically align face images. Experiment results showed that the partially automatic algorithms always performed better than the fully automatic algorithms.

Many of eye detection and tracking algorithms has been proposed in recent years. Current eye detection methods can be divided into two categories: active and passive

eye detection [75]. Passive methods directly detect eyes from images within visual spectrum and normal illumination. The active detection methods use special types of illumination. Under near infrared (NIR) illumination, pupils show physical properties which can be utilized to localize eyes. As a result, in indoor environment and relatively controlled lighting conditions, the passive method achieves very accurate and robust results of eye positions. However, it is also important to pursue visual spectrum methods for the two reasons [76]:

- I. NIR based methods tend to produce a large number of false positives when used in outdoor environment, where the nature light impacts the near infrared illumination.
- II. NIR based methods contribute little to further understand about the human perceptual problem that the brain solves when processing faces in natural conditions.

Hence in the following parts, we will only focus on the passive eye detection methods.

Besides eyes, other facial components (i.e. mouth, eyebrow, and nose) are also important in face-related systems for two reasons:

- I. Only eye positions can not give enough information to draw face boundary and estimate face pose.
- II. In some case, eyes are easily covered by glasses or sunglasses. By detecting other facial components, eye positions can be estimated by other component's positions.

In the following section, the facial component localization is reviewed by classifying into two sub-classes: the selected features representation and the machine learning methods.

2.2.1 The Selected Features Representation

In [35], the facial components were detected using the Haar Wavelet like features set. However, although the Haar features vary with different patterns, sizes and positions, they can only represent the regular rectangular shapes. It was found by the authors that there were insufficient local features to reliably locate each facial component. For components that exhibit large variation in location and appearance on the face, such as the corners of the mouth, there were many false positive candidates. Hence, a global shape constraint was presented and it was reported that the performance of detection was significantly improved.

Gabor filters are popularly used for facial components localization since they are relatively less sensitive to the small shift and deformation. Moreno. et. al. presented an automatic feature selection method that can be applied to different image regions [77]. The representation was based on Gabor features and the methodology automatically selected a set of parameters that were good descriptors for a particular image pattern. The author showed some tests to illustrate the rotation and scale robustness characteristics of the method.

In [78], Feris. et. al. presented an effective object representation using a

constellation of 2D Gabor wavelets that were specifically chosen to reflect the object properties, called as Gabor wavelet networks (GWNs). In the application to facial components localization, the system had two level GWN representations. The first level GWN represented the entire face, which was used to find the face candidates. The second level GWN represented each facial feature, which was initialized in positions according to the affine transformation from the first level GWN.

Smeraldi et. al. proposed a method for facial components detection using a vector of Gabor filters [79]. Two levels of modeling were employed on each facial component. The coarser, or local, model is obtained for each component to extract the Gabor responses at the location of that component in the image of the training set. The finer, or extended model, is obtained by placing the whole retina (point set) at the location of the facial landmark on the training image and collect the Gabor responses from all of the retinal points. All these Gabor features are then arranged into a single vector.

2.2.2 Machine Learning Methods

In [78], the proposed system performed a brute-force search within a limited window for a position that minimizes the score in wavelet subspace between a candidate level-two feature GWN, and the target image. In [79], SVM was employed as the classifiers for the facial components representation. Ma. et. al. [116] used standard Adaboost training methods combined with Viola's cascade approach to build

appearance-based eye detector.

In [80], a fully automatic facial feature point detection system was presented by using Gabor features based on Boosted Classifiers. GentleBoost feature templates were learned using a representative set of positive and negative sample images. It was reported that the GentleBoost gives less emphasis to misclassified examples and converged faster than Adaboost. In the testing phase, the proposed system scanned the image with a 13 by 13 pixels sliding window. Each sliding window had 49 representations (grayscale plus 48 Gabor filters representations). GentleBoost classified outputted a response to measure the similarity between the 49-dimensional representations of the sliding window compared to the learned feature point model. After the scanning, the position with the highest output reveals the feature point position.

2.3 Facial Expression Recognition

A number of successful systems have appeared in the computer vision field for automatic facial expression recognition. Pantic and Rothcrantz [1] identified three basic problems that a facial expression analysis needs to deal with: face detection in a facial image or image sequence, facial expression data extraction and facial expression classification. The review of face detection has been presented in section 2.1. Here we only introduce some FER algorithms according to the later two problems raised by Pantic. More detailed review about the FER systems can be found in

[1][81][90][121][122].

2.3.1 Facial Expression Data Extraction

In [1], Pantic et. al. defined two terms for the facial expression data, named as: facial features and face model features. The facial features were defined as the prominent features of the face, such as eyebrows, eyes, nose, mouth and chin. The face model features were defined as the features used to represent (model) the face. Previously section 2.2 already gave a review for the facial features localization. Hence, in the following part we concentrate on the face model features representation based on the results of face detection and facial component localization.

Michel. et. al. presented a real time approach to emotion recognition through facial expression in live video [82]. An automatic facial feature tracker was employed to perform face localization and feature extraction. The facial feature displacements in the video stream were used as face model features to be input to a Support Vector Machine classifier. The authors mentioned some advantages of this kind of face model features: fast processing speed due to the simple computational complexity, high accuracy on their earlier experimental results and short training and classification delays.

Gabor features are popularly used in some FER approaches [83][84]. A system was built in [83] to automatically detect the frontal face images and classified them into 7 basic facial expressions: neutral, anger, disgust, fear, joy, sadness, surprise. A

Gabor representation of the patch is formed and then processed by a bank of Support Vector Machine classifiers. The face images were converted into a Gabor magnitude representation, using a bank of Gabor filters at 8 orientations and 5 spatial frequencies. After that, the Gabor representation will be sent to an Adaboost classifier for further processing.

Kapoor. et. al. presented a fully automatic framework [85] to analyze facial action units, the fundamental building blocks of facial expression enumerated in Paul Ekman's Facial Action Coding System (FACS). The system employed an infrared sensitive camera equipped with infrared LEDs for pupil detection. Regions of interest corresponding to eyes and eyebrows were cropped out and analyzed according to the detected pupil positions. Principal component analysis (PCA) can be used to figure out the representation of the test image in terms of the linear combination of example images. The experiments conducted by the authors showed that the first 40 eigenvectors sufficed for the task. This projection approach worked particularly well on the subjects who had their images in the training database.

In [86][87], Pantic et. al. introduced two systems for machine recognition of facial muscle actions (i.e., Action Units, Aus) in face video. The first one is the Motion History Images (MHI), which in contrast to Motion Energy Images preserves not only spatial but also the temporal information of the image sequence. Moreover, the value of each pixel in an MHI image indicates where and when motion in the input image sequence has occurred. Higher intensity pixel denotes the recent motion while lower intensity pixel denotes a motion that occurred at an earlier time. After the MHI

calculation, a combined kNN/rule-based classifier was employed for automatic AU detection. The second one is using the tracking technology. 20 fiducial points were localized in the first frame of the input face image sequence. Window around each of the facial points defined a number of color templates. The authors found that particle filtering was a dominant tracking method due to its ability to deal successfully with noise, occlusion and clutter. Hence, it was used for the problem of color-based template tracking. Based on the changes in the positions of fiducial points, the facial expression changes were measured.

W. Zheng et. al. [88] manually located 34 landmark points from each facial image and then converted these geometric points into a labeled graph (LG) vector using the Gabor wavelet transformation method to represent the facial features. On the other hand, for each training facial image, the semantic ratings describing the basic expressions were combined into a six-dimensional semantic expression vector. Learning the correlation between the LG vector and the semantic expression vector was performed by the Kernel Canonical Correlation Analysis (KCCA). According to this correlation, the associated semantic expression vector of a given test image can be estimated.

Moreover, Yeasin et. al. [89] used a linear classification bank on projected optical flow vectors as the facial expression features. Aleksic et. al. [19] utilized facial animation parameters (FAPs), which is supported by the MPEG-4 standard, as the features for classification. Wu et. al. [91] used Active Appearance Models (AAM) to extract 53 facial key points. A system is designed to calculate the shape feature vector

classification by those key points' movements. In [92], Yeasin et. al. employed three different data projection techniques, namely, Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) and Local Linear Embedding (LLE), to project the facial motion onto lower dimensional subspace. Experiments were conducted for comparison of these three methods.

2.3.2 Facial Expression Classification

Cohen et. al. employed Bayesian network classifier for facial expression recognition by using the maximum likelihood estimation [93]. A real time facial expression recognition system was developed based on this algorithm. The system used a model based non-rigid face tracking algorithm to extract motion features input to a Bayesian network classifier, which was used for recognizing facial expressions. The authors claimed that the Bayesian network classifiers were able to handle missing data, both during inference and training. They introduced a stochastic structure search algorithm for learning the structure of Bayesian network classifiers. It was shown that with moderate size labeled training sets and large amount of unlabeled data, this method can utilize unlabeled data to improve classification performance.

Support Vector Machine has been widely used in FER area [82][85]. In [82], Michel and Kaliouby presented an approach based on Support Vector Machine (SVM) for expression recognition in live video. A real time facial feature tracker was used for face localization and feature extraction. They calculate the displacements for each

feature between a neutral and a representative frame of an expression. These are used together with the label of the expression as SVM classifier training input. The trained SVM model is subsequently used to classify the facial feature displacements in real time. The result indicated that the properties of a SVM learning system correlate well with the constraints placed on both accuracy and speed.

As proven to be a successful method in speech recognition, Hidden Markov Models (HMMs) is also a popular tool for FER solution. In [89], HMMs was used to learn the underlying model for the six universal facial expressions. It was trained to recognize the temporal signatures associated with each of the universal expressions. In Aleksic's work [19], eye brow, out-pip or joint facial animation parameters (FAP) vectors, which were extracted from a face image, were used as observations. A HMMs model was used to model temporal changes of facial features. The stream weights were determined based on the facial expression recognition results obtained when FAP streams were utilized individually.

In [91], Y. Wu et. al. presented a method of modeling facial expression space for facial expression recognition by fuzzy integral. In [94], Kim et. al. proposed to use Adaboost to select weak classifiers from a modified rectangular feature set to recognize human facial expression in real time. Bartlett used a combination of Adaboost and SVM to enhance the FER system performance [83].

2.4 Implementation Tools

In this thesis, Fourier Transform is employed to change an input facial expression sequence from the Spatio-Temporal domain into the Spatio-frequency domain. Fuzzy C Means is used as a classifier by giving intermediate memberships of subclasses to represent different expressions. Hausdorff Distance is used to search an optimal result in a small neighborhood. A brief description of these concepts is given in the following sections.

2.4.1 Fourier Transform

Linear transforms are widely used in solving problems in science and engineering. As a pervasive and versatile tool, Fourier transform [95] is frequently used in many fields of science (i.e. mathematical or physical area) to alter a problem into one that can be more easily solved. The Fourier transform is used in digital image processing, linear systems analysis, antenna studies, optics, random process modeling, probability theory, quantum physics, and boundary value problems. Some scientists understand Fourier theory as a physical phenomenon, not simply as a mathematical transform. Generally, Fourier Transform can be seen as a tool that converts a signal into a sum of sinusoids of different frequencies, amplitudes and phases. Fig. 2-1 illustrates the relationship between the time and frequency domain converted by Fourier Transform.

Fig. 2-1 Relationship between time and frequency domain.

2.4.2 Fuzzy C Means

Fuzzy set theory provides a formalism in which the conventional binary logic (hard k means) [96] based on choices "yes (complete true)" and "no (complete false)"

is replaced with a continuum of possibilities that effectively embody the alternative "maybe", which represents the degree of possibility of an occurrence [97]. Fig. 2-2 shows the basic concept of this theory. Fuzzy C Means (*FCM*) is one useful approach for fuzzy classification, which can determine the intrinsic division in a set of unlabeled data and find representatives for homogeneous groups. Unlike hard C-means algorithm which executes a sharp classification in which each object is either assigned to a class or not, *FCM* can make an object belong to several classes at the same time but with different degrees between 0 and 1. Since some expressions might look alike at certain facial regions, it is hard to put clear boundaries between different expressions. Thus, *FCM* is used since it can give intermediate memberships instead of hard boundaries.

Fig. 2-2 Examples of Non Fuzzy set and Fuzzy set.

2.4.3 Hausdorff Distance

Hausdorff Distance (HD) measures how far two compact subsets of a metric space are from each other. Because of its non-explicit feature (point) correspondence and computational simplicity, many researchers further investigated HD to enhance its performance [98] and applied HD on different practical applications. The most interesting works include face recognition [99][100][101], word recognition [102] and video sequence matching [103].

Given two sets of points $A = a_1, \dots, a_m$ (representing a model in the database) and $B = b_1, \dots, b_m$ (representing a test image), HD was defined as

$$H(A, B) = \max(h(A, B), h(B, A))$$

with

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

where $\|\bullet\|$ is any norm, e.g. Euclidean Distance. The function $h(A, B)$ is called the directed Hausdorff Distance from A to B . It identifies the point $a \in A$ that is the farthest from any point of B , and returns the distance from such a point to its nearest neighbor in B . Thus $H(A, B)$ measures the degree of mismatch between two sets, as it reflects the distance of the point of A that is the farthest from any point of B and vice versa. Intuitively, if $H(A, B)$ is d , then every point of A must be within a distance d of some point of B and vice versa [61].

2.5 Summary

This chapter presents an overview of face detection, facial component localization and facial expression recognition approaches. It also gives a brief introduction about Fourier transform, Fuzzy C Means and Hausdorff distance. By dividing the FER approaches into two main problems: facial expression data extraction and facial expression classification, some typical methods are analyzed and reviewed. Inspired by these approaches, a FER system has been built and will be introduced in the following sections, which aims to solve some current existing problems and challenges. Experiments analysis of the system performance will also be shown in the later part of the thesis.

Chapter 3 Pre-processing

To alleviate the influence of variation of pose and face size, each input face image needs to be normalized before further processing. Rough face detection results are not accurate enough for face normalization. Here we adopted Viola's system for face detection and proposed a novel feature descriptor to localization the precise position of each facial component (i.e. left and right eye corners). The flowchart of the pre-processing system is displayed in Fig. 3-1.

Fig. 3-1 The system pre-processing flowchart.

3.1 Input Data Image

The system accepts a sequence of images displaying a particular expression as the input data. This is based on the psychological study in [42] that temporal information can reveal more relevant information about the underlying emotional states than a single static image. The Cohn-Kanade and MMI facial expression databases are used as the sources of the input sequences.

3.1.1 Cohn-Kanade Database

The Cohn-Kanade database is one of the most comprehensive test-beds to date for comparative studies of facial expression analysis [43]. It consists of 96 subjects with 391 expression sequences. To be comprehensive, this database has included the variation of different age, gender and race. The subjects who played the facial expression ranged from 18 to 50 years old, including male, female, Euro-American, Afro-American, and other groups of people. The subjects were observed in an observation room equipped with a chair on which to sit and two cameras, each connected to a video recorder with a synchronized time-code generator. Both the frontal view and 30 degrees profile view were recorded, but only the image sequences from the frontal view camera are experimented in this work.

The frontal view image sequences were digitized into either 640x490 or 640x480 pixel arrays with 8-bit gray value or 24-bit color values. To be tested in a uniformed standard, a sub image with resolution of 200x220 has been extracted and normalized

from each original image. The extracted contains the main facial area and excludes the useless background. The images with the 8-bit gray value format are used in this work. One example from the database is displayed in Fig. 3-2.

Fig. 3-2 An example from the Cohn-Kanade database.

(a) full image; (b) main facial area.

3.1.2 MMI Face Database

The MMI Face Database is an easily accessible, comprehensive, benchmark database of facial expression [114]. It contains more than 1500 samples of both static images and image sequences of faces both in frontal and in profile view, displaying

various facial expressions of emotion, single AU activation, and multiple AU activation. Here we only concentrate on the image sequences part.

In the MMI face database, all video sequences have been recorded at a rate of 24 frames per second using a standard PAL camera. The database contains approximately 30 profile-view and 750 dual-view facial expression video sequences. These videos are captured from 19 different faces of students and research staff members of both sexes (44% female), ranging in age from 19 to 62, including either a European, Asian, or South American ethnic background.

A sub image with resolution of 200x220 (same with the Cohn-Kanade database) has been extracted and normalized from each original image.

3.2 Pre-processing

3.2.1 Face Detection

As mentioned in chapter 2, Viola's work has achieved great success in face detection with fast processing speed and high detection accuracy. OpenCV (Open Source Computer Vision Library) [104] has included the statistical approach originally developed by Viola and Jones [34] and then later analyzed and extended by Lienhart [67] for face detection. OpenCV provides low-level and high-level APIs for face detection. The low-level API allows users to check an individual localization within the image by using the classifier cascade to find whether it contains a face or not. Helper functions calculate integral images and scale the cascade to a different

face size (by scaling the coordinates of all rectangles of Haar-like features). Alternatively, the higher-level function does this all automatically, and it can achieve good enough results in most cases [105]. Here we adopt the low-level function for our proposed system. 8000 training face images and millions training non-face images were fed to the system to calculate the trained parameters.

3.2.2 Facial Component Localization

As mentioned earlier, the gross position of face is not good enough to satisfy the requirement for further face processing. Precise facial component localizations (normally eye positions) are very important for face normalization. Some researchers observed that Viola's system achieved great success in face detection, so they attempted to use the similar system to locate facial components [78], employing the Haar features. The results are not good enough since Haar features have limited discriminant capability. Although in some systems the Haar features vary with different patterns, sizes and positions, they are best to represent the regular rectangular shapes [36]. However, for the facial component localization, the most distinguishing features are some curve shapes that are quite hard to be represented by rectangles.

In face detection area, Adaboost has been proven to be a powerful tool to select the most discriminated features from a feature bank. Also, we observed that Gabor filter is good at feature representation and gives good results in the face recognition

and facial component localization. Motivated by this, we propose a new feature to represent the curve shapes of each facial component by using Gabor filter. The method is described in the following parts of this chapter.

3.2.2.1 System Overview

We propose a method to precisely determine facial component localizations (9 pixels) from a face detector output. Fig. 3-3 displays the 9 facial component points. They are the 4 inner and outer eye corners, 2 eyebrow tips, 2 mouth corners and nose. Unlike some systems detecting the centers of pupils, we locate the precise positions of the eye inner and outer corners. From our observation, it is found that the pupil moves inside the eyelids with variable shape, size and color. The eye corner shapes are relatively more stable. Hence, the positions of two eye corners (inner and outer) are found more suitable than the pupils to compute an invariant distance to normalize the input face.

We use boosting with linear classifier as weak hypothesis and rectified Gabor filters as features to locate 9 facial components positions. Locating 9 features instead of only two eyes has some advantages. Sometimes the eye may be covered by glasses or sunglasses while mouth sometimes may be covered by the moustache or beard. In these cases, other facial components can help the system to estimate the hidden facial component positions. Some examples of occluded facial component localization have been displayed in the Appendix I.

Fig. 3-3 The 9 facial components for normalization.

3.2.2.2 Training Samples

We generate the positive and negative training samples from human face images for the Adaboost. 9 facial components are marked on the training images manually. The marked point is a positive training sample for the corresponding facial component, while other points except this are the negative training samples. Since the face is symmetry, for each facial component, each face can offer two positive samples. Fig. 3-4 further shows some examples of the eye training data.

Fig. 3-4 Some examples of eye inner corners positive training samples.

3.2.2.3 Feature Extraction

We first define Gabor filter bank $G \{g_0 \dots g_{11}\}$ in Fig. 3-5 and define the list of point sets $P \{p_0 \dots p_5\}$ in Fig. 3-6. Each point set p_k has 5 pixels $pp_k = \{pp_{k0} \dots pp_{k4}\}$. Hence, point set p_k can be represented by a 60-D vector V (12 Gabor filters $\times$ 5 points). A scan window W (the white box in Fig. 3-5 as an example) can calculate 6 vectors, corresponding to the 6 point sets. At each stage, Adaboost will choose the most

discriminable V from the 6 point sets as the weak classifiers of that stage.

Fig. 3-5 Left: one point set p_k on the face; Right: Gabor filter bank $G \{g_0 \dots g_{11}\}$.

Fig. 3-6 Six point sets for Adaboost selection $\{p_0 \dots p_5\}$.

Given a scan window W , the vector $V_{P,G}$ is defined as:

$$V_{P,G}(W) = \Lambda * [v_{ij}]^T \quad \text{where } i=0 \dots 4 \text{ and } j=0 \dots 11$$

$$v_{ij} = W(pp_i) \circ g_j \quad (3.1)$$

$$\Lambda = \frac{1}{\sum_i \sum_j |v_{ij}|}$$

In equation 3.1, “ $\circ$ ” denotes the dot product (convolution) operator and v_{ij} is also

known as the response of Gabor filter g_j with W at location pp_i . The feature vector is finally L1-normalized by using Λ . The dimension of V_{PG} is $\dim = 5 \times 12 = 60$.

The reason we choose Gabor feature for localization instead of Haar feature is that Gabor filters can give more precise information about the curve/shape of the facial component. Because of their multiplication-convolution property, Gabor filters are widely used in image processing application, such as iris recognition, fingerprint recognition and facial component detection [77][78][79][80]. The promising experiment results shown in Chapter 5 also prove that our proposed Gabor features are more suitable for facial component localization problems.

3.2.2.4 Classifiers Training Procedures

For the 9 facial component feature points, the localization procedures are the same. Here we only use the left eye inner corner as an example.

We adopt Viola's cascade structure to build our Adaboost classifier. Each Adaboost classifier for facial component localization consists of several strong classifiers, while each strong classifier consists of several weak classifiers. Given a scan window W , it can be described as six 60-d vectors (as discussed in section 3.2.2.3). To calculate a weak classifier, the parameters trained by Adaboost project these 6 vectors into a 1-d space, generate 6 scale scores and select the vector with the maximum score as the weak classifier. The maximum score reflects the possibility that scan window W contains the left eye inner corner for the corresponding weak

classifier.

If 2 facial components are symmetrical (i.e. left & right eye inner corners), we use the same 60-d parameters and simply reverse the image on the x axis. The left image of Fig. 3-7 displays some output scores of the corresponding scan window for one 60-d to 1-d projection. The right image is exactly the reversed images on x axis of the left one.

Fig. 3-7 Examples of left/right inner eye corner projected scores.

Left: the different scores for corresponding scan windows; Right: the flipped image to locate the symmetrical facial components.

To estimate the 60-d parameters, we employ Mp positive training samples and Mn negative training samples. The system will generate an output of 1 or -1 according to the positive or negative sample respectively. The parameters need to satisfy the condition: $V \circ B = O$. The parameters V , B , and O are defined in Fig. 3-8.

Fig. 3-8 The equation for parameters (B) estimation.

In real application, it is very hard to find such a solution for the equation. Alternatively, we estimate B as: $B = \arg \min_w w \times |V \otimes B - O|$, where w is the self updated weight value. The training parameter setting follows the work done in [106].

The training procedures are described as the following.

INPUT:

1. N positive eye training samples $\{W, y=+1\}$.
2. M negative non-eye training samples $\{W, y=-1\}$.
3. Gabor Filter Bank $G=\{g_0 \dots g_{11}\}$ with 12 filters.
4. A list of 6 point sets $P = \{p_0 \dots p_5\}$, where each point set has 5 points $pp_k = \{pp_{k0} \dots pp_{k4}\}$
5. The previous strong classifier F_{n-1} . (For training F_0 , use $F_{-1}(W)=0$).
6. Minimum hit rate $HR_{\min}$ and maximum false positive rate $FPR_{\max}$ of strong classifier. We use $HR_{\min}=0.99$, $FPR_{\max}=0.30$.

INITIALISE:

1. Initial strong classifier $F(W)=0$.

TRAIN STRONG CLASSIFIER

1. Set the weight ω of each training sample as: $\omega(W) = \exp(-y * (F(W) + F_{n-1}(W)))$.
2. Normalize the weight such that $\sum \omega = 1.0$
3. For each of the point set P_k ($k=0 \dots 5$), build a weak classifier $f(z)$ using steps 3.1. to 3.4.
 - 3.1 For each sample W , extraction the 60-d feature vector V of point set P_k .
 - 3.2 Use weighted least square method [107] to find the best linear separating hyper plane $V \circ B = O$, where $O = \{\pm 1\}$.
 - 3.3 Find the projection values z of the feature vectors V onto B using $z = V \circ B$. Build a histogram $f(z)$ to estimate the conditional probabilities.

$$f(z) = 0.5 * \log \left(\frac{\sum \omega_z^+ + \delta}{\sum \omega_z^- + \delta} \right)$$

Here ω_z^+ and ω_z^- are the weights of the positive and negative samples with value z respectively. δ is a small positive constant to prevent numerical instability. We use number of bins of histogram, $|f|=16$ and set $\delta=0.05/|f|$.

- 3.4 Estimate the classification error ε of $f(z)$ as:

$$\varepsilon = 0.5 * \sum \varepsilon_z \quad \text{where} \quad \varepsilon_z = \sqrt{\omega_z^+ * \omega_z^-}$$

4. Select the point set P_{selected} that gives the $f(z)$ with the smallest error ε from P_k ($k=0 \dots 5$). Update the current strong classifier as:

$$F(W) = F(W) + f_{\Theta, \text{selected}}(W) \quad \text{where } \Theta = \{ P_{\text{selected}}, G, B_{\text{selected}} \}$$
5. Using the current $F(W)$, determine the current false positive rate FPR at hit rate $HR = HR_{\min}$. If $FPR < FPR_{\max}$, exit. Else go to step 1.

OUTPUT

1. Trained strong classifier: $F(W) = \sum f_{\Theta, \text{selected}}(W)$.

3.2.2.5 Facial Components Localization Procedures

Fig. 3-9 shows the structure of the Adaboost facial components locator system.

Fig. 3-9 Facial component localization system.

Given an input face image, the multi-scale representation (image pyramid), is constructed. The number of scale for image pyramid is dependent on the accuracy and speed requirement. More scales used, more confidential results you will get and the system needs more processing time. The image pyramid is shown in Fig. 3-10.

Fig. 3-10 An example of pyramid images.

The scanner extracts a scan window measuring 48×48 , running from top-to-bottom, left-to-right for each of the multi-scale images, within a predefined search region. For each scan window W , Adaboost classifier would give a score H (displayed in the left image of Fig. 3-7) to indicate the probability likelihood ratio of

the presence of an eye at the center of W . The rough procedures are described as the followings.

INPUT:

1. A list of all M scan window in search region, $L_{ALL} = \{W_0, W_1, \dots, W_{M-1}\}$.
2. A cascade of N strong classifiers $\{F_0, F_1, \dots, F_{N-1}\}$.

INITIALISE:

Set scan window list, $L = L_{ALL}$. Size of L is $|L|=M$.

RUN CASACDE CLASSIFIER

For $m=0, \dots, |L|-1$

For $n=0, \dots, N-1$,

Run strong classifier F_n on W_m .

If $F_n(W_m) < \text{Threshold}_n$, remove W_m from L . Threshold_n is the reject threshold of strong classifier F_n .

End

End

OUTPUT

A list of remaining scan windows in L .

After all scanning windows are processed, the final localization (x, y) is calculated as:

$$\begin{aligned} x &= \Omega * \sum_m^{|L|} F(W_m) * x_m & y &= \Omega * \sum_m^{|L|} F(W_m) * y_m \\ \Omega &= \frac{1}{\sum_m^{|L|} F(W_m)} \end{aligned} \quad (3.2)$$

In above equation, L is the list of remaining scan windows W_m after the cascade process. $|L|$ is the size of L , i.e. number of remaining scan windows. $F(W_m)$ is the classifier score when W_m is pass through the cascade. (x_m, y_m) is the center of the window W_m . W_m must be scaled back to the normalized image. Ω is the normalizing

constant.

3.2.2.6 Facial Component Localization on Occluded Face

Sometimes the input face images may be occluded and some facial components are covered by other external objects, i.e. glasses, hand or mask. In such cases, the blocked facial component's position needs to be predicted by other detected feature positions. In [115], Cootes et al. argued that a known rigid object should only be able to deform in characteristic ways when it is in the presence of noise, clutter and occlusion. Cootes employed Principle Component Analysis (PCA) to constrain the global shape variation. Here we use the similar way to calculate the occluded facial component positions.

PCA Parameters Training

For each face, it contains 9 facial components. These components can be represented by a vector x of 18 dimensional. Given a training set of manually localized faces, the mean shape is calculated using:

$$\bar{x} = \frac{1}{N_{set}} \sum_{i=1}^{N_{set}} x_i \quad \text{where } N_{set} \text{ is the number of faces} \quad (3.3)$$

For each face in the training set, its deviation from the mean, dx_i , is calculated as:

$$dx_i = x_i - \bar{x} \quad (3.4)$$

We can then calculate the 18×18 covariance matrix S as:

$$S = \frac{1}{N_{set}} \sum_{i=1}^{N_{set}} dx_i dx_i^T \tag{3.5}$$

The eigenvectors p_k and eigenvalues λ_k are calculated to satisfy:

$$Sp_k = \lambda_k p_k \quad \text{where } \lambda_k \geq \lambda_{k+1}$$
$$\lambda_T = \sum_{i=1}^{N_{set}} \lambda_i \tag{3.6}$$

The first 5 eigenvalues are displayed in Table 3-1. In our system, only these 5 eigenvalues and the corresponding 5 eigenvectors are used to predict the occluded facial components. Other eigenvalues are ignored since they are too small. We combine the corresponding 5 eigenvectors into a Matrix and call it as $P=(p_1, p_2, ..., p_5)$.

Table 3-1 Eigenvalues driven from a set of face images.

Eigenvalue	$\frac{\lambda_i}{\lambda_T} \times 100\%$
λ_1	0.553
λ_2	0.282
λ_3	0.028
λ_4	0.026
λ_5	0.018

Occluded Facial Components Prediction

When we use the algorithm described in section 3.2.2.5 to localize the facial components, each facial component will get a classifier score $\sum_m^{|L|} F(W_m)$. If this score is larger than the corresponding threshold, this component is accepted to have been localized successfully. It will be placed into a set R_{suc} . Otherwise, this component is assumed to have been occluded. It will be placed into the set R_{occ} . The occluded facial

components are predicted in the following step:

1. Set vector $b=0$.
2. Calculate all the facial components localization displacement as:
 $dx_i = x_i - \bar{x}$. If the i^{th} position is in the set R_{occ} , set $dx_i = 0$.
3. Referring to [115], calculate the PCA parameters displacement as:
 $db = P^T dx$.
4. Update the facial components positions as: $x = \bar{x} + P(b + db)$.
5. If the dx is smaller than a threshold T_d or the iteration number is larger than the threshold T_n , break; otherwise, go to step 2.
6. $x = \bar{x} + P(b + db)$ is the new predicted facial components positions.

3.2.2.7 Discussion

Our proposed algorithm about facial components localization is inspired by [68] and [80]. In [68], Abramson. et. al. proposed a system based on sampling individual pixels within the scan window for face detection. In [80], Gabor features were used for the same problem and good performance was reported. The system in [80] used 8281 features to represent one point, which is the center of a 13×13 pixels image patch. We found that 13×13 is a relatively small region, which may concentrates on the local features too much and ignores some global patterns of facial components. In our proposed system, the scan window is 48×48 . Good performance, both in accuracy and speed, will be reported in Chapter 5.

3.3 Facial Area Extraction & Normalization

For facial expression recognition, usually the input images include not only the facial area, but also the unwanted background, which may complicate the recognition process and confuse the system. To avoid this kind of irrelevant information, there is a need to extract and normalize the facial area before the recognition process. Moreover, in real scenarios, due to different distances between people and camera, the size and pose of the face vary unpredictably. To make the system more robust and reliable, the proposed system needs to handle different resolutions of facial area on the input images. When an unknown sequence is presented, the system should normalize the input images with respect to the distances between two eye centers of a face. We set D as the standard distance between two eyes, and d_k the actual distance between two eyes at the k^{th} frame of the input sequence. Thus, the k^{th} frame can be normalized (scaled) by D/d_k . Furthermore, it is quite easy to locate and normalize the facial area if the eyes positions are available as shown in Fig. 3-11, where one example of facial area normalization based on the eye positions is displayed. On the normalized image, a fixed-size rectangle is extracted and fed to the next step.

Currently, we only use the eye corners for face normalization. The reason is that other facial components, i.e. eyebrow and mouth corners, change dramatically when people play the expression. Other facial components are used when the eye corners are occluded. In future, we may use all the 9 components to estimate the depth information, which will be discussed in chapter 6.

Fig. 3-11 An example of facial area normalization.

Left: the original image with detected face and 9 located facial components; Middle: the normalized face; Right: the main face area.

3.4 Frame Selection

Frame number is another important parameter of the proposed FER system. It is required that one input sequence contains only one iteration of facial expression movement, i.e. from neutral state with no expression to maximum muscle distortion. If the sequence has too many frames, the difference between two consecutive frames will be quite imperceptible. Furthermore, since the system needs to handle all the input frames, the redundant and overlap information may slow down the computation and mislead the system judgment. On the other hand, if the sequence includes not enough frames, the movements between two consecutive frames may be dramatic and some details may be lost. The missing information may influence the system performance.

In this study, we have tested the system performance with different numbers of frames in a sequence. The optimal result and parameters setting will be described in

the latter part. Here, we discuss the frame selection process to prepare and select input sequences for the system.

Frame Selection Procedure

1. Assume that there are Z frames in the original sequence but M frames should be kept.

If $Z < M$, this sequence will be ignored; else if $Z = M$, this sequence is accepted. In either case, we will not execute step 2 to 6. Within the facial area, the intensity value at each pixel position can be represented as $f(i, j, t)$, where i and j specify the localization while t is the timing parameter.

2. In current setting, we have $M=11$ and will keep the first and the last frames.
3. Let Z' be the current number of frames in the sequence. If $Z' = M$, it is done.
4. For each t^{th} ($2 \leq t \leq Z' - 2$) frame, calculate the difference picture as

$$Difference(t) = \sum_i \sum_j |f(i, j, t+1) - f(i, j, t)|$$

5. Select the minimal $Difference(t)$ and delete the corresponding t^{th} frame.
6. Repeat step 3 to 6.

3.5 Summary

This chapter describes a facial component localization method using Gabor filter for feature representation and Adaboost for selecting the most discriminated features. This facial components localization (especially eye localization) is an essential preprocessing step for further face processing. It helps to normalize the face area

image to alleviate the influence of variation of different pose and face size. Following this, we will introduce our facial expression recognition system in next chapter.

Chapter 4 Facial Expression Recognition

A facial expression recognition (FER) scheme will be described in this chapter. The proposed system accepts sequences of images as input data of facial expression. This sequence displays one iteration of a particular facial expression, changing from neutral state (no emotion) to extreme muscle distortion. Prior analyzing the input data, the pre-processes displayed in Fig 3-1 have to be first applied on the input. These include Eye Detection, Facial Area Extraction, and Frame Selection.

In [64], two essential issues are mentioned from general pattern recognition viewpoint: feature selection and classifier design based on the selected features. For our proposed facial feature recognition system, we have derived novel local feature descriptors based on Fourier Transform to obtain discriminative feature. After that, the Wave shape and Fuzzy C means (FCM) are employed to recognize facial expression. Our work started from 1D Fourier Transform and extended to 2D Fourier Transform to include both temporal and spatio information in the descriptor. We find that the transformed frequency domain representation of the original Temporal-Spatio information is accurate and effective for the FER problem.

Fig. 4-1 displays the basic structure of our system. This chapter is organized as the following: section 4.1 introduces our feature descriptor representation, section 4.2

illustrates the employed classifiers, section 4.3 gives the detail information of building the expression model and section 4.4 shows the steps to do expression recognition.

Fig. 4-1 Structure of the FER system.

4.1 Feature Descriptor Representation

4.1.1 1D Fourier Transform Extraction

The facial area can be extracted by the pre-processes described in Chapter 3. Within the facial area, the intensity value at each pixel position can be represented as $f(x, y, t)$ as shown in Fig. 4-2, where x and y specify the localization while t is the timing parameter with $1 \leq t \leq M$, $1 \leq x \leq X$, $1 \leq y \leq Y$.

Fig. 4-2 An expression sequence seen in 3D space.

Since the Fourier Component spectrum in the frequency domain can indicate the approximate shape of the original wave on the temporal axis, the 1D Fourier Transform is employed to transform the input gray value intensity wave form into frequency domain. From Fourier theorem, the Component zero only reflects the sum of gray value intensity and is independent with the variation on the temporal axis. Thus the Component zero is ignored by the system. Fig. 4-3 and 4-4 show an example of input and output data of 1D Fourier Transform of a particular pixel on the temporal axis. Since M frames will be analyzed, each pixel in the facial area gets M gray level intensity values ($0 \leq k \leq M-1$). The 1D Discrete Fourier Transform can be computed on each pixel as

$$F(x, y, m) = \frac{1}{M} \sum_{k=1}^M f(x, y, t) e^{-jt2\pi m / M} \quad (4.1)$$

Since $e^{-j2\pi tm/M} = \cos(2\pi tm/M) + j \sin(2\pi tm/M)$, function (4.1) can be written as

$$F(x, y, m) = \frac{1}{M} \sum_{t=1}^M (f(x, y, t) \cos(2\pi tm/M) + j f(x, y, t) \sin(2\pi tm/M)) \quad (4.2)$$

The Fourier Transform can be described by two parts: the real and imaginary parts, which satisfy equation (4.3), (4.4), and (4.5).

$$F(x, y, m) = \text{real}(x, y, m) + j \text{img}(x, y, m) \quad (4.3)$$

$$\text{real}(x, y, m) = \frac{1}{M} \sum_{t=1}^M f(x, y, t) \cos(2\pi tm/M) \quad (4.4)$$

$$\text{img}(x, y, m) = \frac{1}{M} \sum_{k=1}^M f(x, y, t) \sin(2\pi tm/M) \quad (4.5)$$

Fig. 4-3 An example of input data of 1D Fourier Transform.

Fig. 4-4 An example of output data of 1D Fourier Transform.

(the first component has been ignored)

4.1.2 2D Fourier Transform Extraction

The 1D Fourier Transform only utilizes the frequency information on the time axis. It suffers from one limitation that the performance may drop when the time axis information is not enough (i.e. only a few frames are used for Fourier Transform; M is too small). In such case the local feature descriptor have only a few dimensions and can not discriminate different feature well. Inspired by this view, we extended our work from 1D to 2D Fourier Transform.

To emphasize the movement of face muscle, we compute the differences between the first frame and other frames as

$$f'(x, y, t) = f(x, y, t) - f(x, y, 1) \quad (4.6)$$

where $t = 1, 2, \dots, M$.

The collection of all $f'(x, y, t)$ composes a $M \times X \times Y$ 3D space. Some examples of $f'(x, y, t)$ from an 11-frame sequence are displayed in Fig. 4-5. From this 3D space (Fig. 4-5 (g)), one can extract temporal patterns from a fixed location (i.e. the 1D temporal pattern with respect to t , described in previous section), from a horizontal or vertical line (2D pattern with respect to t - x or t - y as shown in Fig. 4-5 (g)), or from a small 2D area (3D pattern with respect to t - x - y). We have done the investigation on the 1D and 2D temporal patterns in this thesis.

Along the x axis, each row of X pixels are divided into N_p pieces here, i.e. each piece has N_{pn} pixels, where $N_{pn} = X / N_p$. Each horizontal piece forms a small 2D sub matrix with respect to time with dimension $M \times N_{pn}$. The $f'(x, y, t)$ falling in this sub matrix is renamed as $f''_{x_0 y}(t, x')$ with $1 \leq t \leq M$ and $1 \leq x' \leq N_{pn}$ as shown in Fig. 4-5 (g) where x_0 is an offset specifying the smallest x value of the line. We apply the 2D Discrete Fourier transform on each sub matrix as

$$F_{x_0 y}(u, v) = \frac{1}{M * N_{pn}} \sum_{t=1}^M \sum_{x'=1}^{N_{pn}} f''_{x_0 y}(t, x') e^{-j2\pi(ut/M + vx'/N_{pn})} \quad (4.7)$$

with “ j ” representing imaginary. For an angle ϑ , we have $e^{-j\vartheta} = \cos(\vartheta) + j\sin(\vartheta)$.

a small 2D sub matrix with dimension $M \times N_{pn}$
(pixels falling in this area is renamed as $f''_{x_0,y}(t, x')$)

Fig. 4-5 Examples of $f'(x, y, t)$ in 3D space.

(a) $f'(x, y, 1)$; (b) $f'(x, y, 3)$; (c) $f'(x, y, 5)$; (d) $f'(x, y, 7)$; (e) $f'(x, y, 9)$;

(f) $f'(x, y, 11)$; (g) collection of $f'(x, y, t)$.

Equation (4.7) can be rewritten as the combination of the real and imaginary parts as

$$F_{x_0,y}(u,v) = \text{real}_{x_0,y}(u,v) + j \times \text{img}_{x_0,y}(u,v)$$

with

$$\text{real}_{x_0,y}(u,v) = \frac{1}{M * N_{pn}} \sum_{t=1}^M \sum_{x'=1}^{N_{pn}} f_{x_0,y}''(t, x') \cos(2\pi(ut / M + vx' / N_{pn}))$$

and

$$\text{img}_{x_0,y}(u,v) = \frac{1}{M * N_{pn}} \sum_{t=1}^M \sum_{x'=1}^{N_{pn}} f_{x_0,y}''(t, x') \sin(2\pi(ut / M + vx' / N_{pn}))$$

Since we have $N_p \times Y$ sub matrices, we will need to repeat the computation $N_p \times Y$ times. Examples of the computed Fourier spectra can be found in Fig. 4-6 (b), (d), (f) and (h).

4.2 Classifiers Processing

An expression can be represented and described by its specific muscle movements at different localizations inside the facial area. For example, the mouth movement is quite dramatic for smile and surprise. Comparatively, it is much mild for the sad expression. Fig. 4-7 shows four examples of mouth movements for four expressions. Consequently, the modeling needs to record both the localization and movement information. In this work, Fourier Components are used to describe the movement while Modified Hausdorff Distance measures the difference of the localization.

Fig. 4-6 Examples of the expression Fourier spectra.

(a) the original image and selected rows; (b) Fourier spectra along the selected rows with $N_p = 5$; (c) the area for training ($N_p = 5$, size= 11×40); (d) Fourier spectra with $N_p = 10$; (e) the area for training ($N_p = 10$, size= 11×20); (f) Fourier spectra with $N_p = 20$ (enlarged 4 times for viewing); (g) the area for training ($N_p = 20$, size= 6×10); (h) Fourier spectra with $N_p = 40$ (enlarged 9 times); (i) training area ($N_p = 20$, size= 4×5).

For each expression, we assume that the number of 1D or 2D Fourier

Component patterns is limited. It is of interest to find such a number, c , and then classify the patterns into c different classes accordingly. First we describe the movement using Fourier Components and make a feature vector $v(x, y)$ (discussed later in section 4.2.2.1). The feature vector $v(x, y)$ is further refined to find the c classes. Finally, each of the 6 expressions (smile, surprise, anger, disgust, fear and sad) is modeled by the membership function of the Classifiers. Details of the construction are described below.

Fig. 4-7 Examples of different types of mouth movements with expressions

(a) neutral, (b) smile, (c) surprise and (d) sad.

4.2.1 Wave Shape Approach

4.2.1.1 Feature Extraction

The relationship between the original wave shape and the Fourier spectrum is described in Fig. 4-8. It is observed that the position of the Fourier Component in the frequency domain indicates the number of peaks in the Spatio-Temporal domain. Hence, the pixels on the facial area can be classified into several classes according to the dominant frequency. Considering there is only one iteration of facial motion in the sequence, the wave shapes with more than 2 peaks are quite rare in our experiments. In practice, we classify the wave shapes into 3 classes. Some typical examples of wave shape and the corresponding Fourier Components are displayed in Fig. 4-8.

The first component ($F(x, y, 1)$) which reflects the lighting intensity level will be ignored while the other components which capture the frequency information of the temporal pattern will be kept to form representation that is relatively insensitive to lighting change. We collect the first N_d Fourier component pairs (the real and imaginary parts) to represent and describe the temporal pattern. The loss information can be computed in the following equation with $m=2, 3, \dots, N_d+1$. We choose a large enough N_d in order to have the *Loss* information proportionally less than 10%.

$$Loss = \sum_{m=N_d+2}^M |F(m)| / \sum_{m=2}^M |F(m)| \quad (4.8)$$

From our observation, it is found that some facial area, i.e. cheek and mouth, exhibit significant muscle movement when facial expression is performed; some other facial area, i.e. forehead and nose, exhibit far less movement. To emphasize the

significant movement, speed up computation and ignore irrelevant and unnecessary information, one can consider those pixels that exhibit significant expression energy only, i.e. each pixel is ranked by

$$Sum(x, y) = \sum_{m=2}^{N_d+1} |real(x, y, m)| + |img(x, y, m)| \quad (4.9)$$

Only the highest $N_{\%}$ pixels are fed to the system for the facial expression model training and recognition in later section. From previous discussion, we have concluded that the dramatically moving area concentrates on the eye and mouth parts of the face. The proportion of this area is approximate 30% of the whole face area. From calculation, we found that the largest 10% of $sum(x, y)$ exhibits approximate 30% motion energy of the whole facial area while the largest 30% of $sum(x, y)$ exhibits approximate 60% motion energy. To avoid expensive computation, these largest 30% $Sum(x, y)$ values form a set R' . However, the area of one peak wave shape is much larger than that of the two peaks wave shape in R' . To pick up the most dominant one peak wave shape, the largest 10% $Sum(x, y)$ values are selected to form another set R for this purpose. The pixels will be classified by equation (4.10) as

$$DFC(x, y) = \begin{cases} 1 & (x, y) \in R_1 \\ 2 & (x, y) \in R_2 \\ 0 & (x, y) \in R - R_1 - R_2 \end{cases} \quad (4.10)$$

where DFC stands for Dominant Fourier Component. R_1 is the set of pixels in R that satisfies conditions 1 and 2 below:

- 1 Amplitude of the first component is the highest.

- 2 Amplitude of the first component is at least 50% larger than other.

R_2 is the set of pixels in R' that satisfies conditions 3 and 4:

- 3 Amplitude of the second component is the highest.
- 4 Amplitude of the second component is at least 20% larger than other.

4.2.1.2 Model Construction & Recognition

With the Wave Shape classification, an expression sequence can be described by three sets of DFCs. If the input images are from a model sequence, the three sets of classification results can be stored and act as model templates. If the input images are from an unknown test sequence, the three sets of classification results will be compared with the model templates from the database. We expect that same expression should display similar motion pattern on the facial area.

To measure the dissimilarity of DFCs and match one expression to another, one can use the MHD (Modify Hausdorff Distance) [56] to compute dissimilarity scores from an input to 6 expression model point sets, i.e. distance from $DFC_{input}(k)$ to $DFC_{model}(k)$ with $(0 \leq k \leq 5)$. Let $N(i)$ stand for the number of pixels in $DFC_{model}(i)$. (x, y) is one of the facial points on the model data while (x', y') is a point on the neighborhood around (x, y) of the input image. The MHD dissimilarity value (DI) between an Input sequence and a Model sequence is computed using the equation 4.11 – 4.15.

Fig. 4-8 Examples of $f(x, y, t)$ and Fourier Components
with (a) one peak; (b) two peaks; (c) mixture of peaks.

$$DV = \frac{\sum_{i=0}^2 \sum_{(x,y) \in DFC_{\text{model}}(i)} D_i(x,y)}{\sum_0^2 N(i)} \quad (4.11)$$

$$D_i(x,y) = \min_{(x',y') \in DFC_{\text{input}}(i)} (D_i'(x,y,x',y') \times (1 + \tan \theta_{(x,y,x',y')}^i)) \quad (4.12)$$

$$\theta_{(x,y,x',y')}^i = \begin{cases} |\theta_{(x',y')}^i - \theta_{(x,y)}^i| & \text{if } i=1,2 \\ 0 & \text{if } i=0 \end{cases} \quad (4.13)$$

$$\theta_{(x',y')}^i = \arctan \frac{\text{img}(i)_{(x',y')}}{\text{real}(i)_{(x',y')}} \quad \text{if } i=1,2 \quad (4.14)$$

$$\theta_{(x,y)}^i = \arctan \frac{\text{img}(i)_{(x,y)}}{\text{real}(i)_{(x,y)}} \quad \text{if } i=1,2 \quad (4.15)$$

where θ is the phase shift, $\text{real}(i)$, $\text{img}(i)$ are the corresponding Fourier Components in Equation (4.3) and $D_i'(x,y,x',y')$ is the L_2 Distance (Euclidean Norm) between (x,y) and (x',y') from the i^{th} class.

Equation 4.12 is employed to add in the phase difference of two pixels from the same DFC. The product of the phase difference and the geometrical difference finally give the MHD dissimilarity scores. The model with the minimum MHD is the match.

4.2.2 Fuzzy C Means Approach

4.2.2.1 Feature Extraction

It is observed that the pixel-based temporal pattern, $f(x, y, t)$, exhibits unique patterns of 1 peak, 2 peaks or a mixtures of them as shown in Fig. 4-8 where three examples of $f(x, y, t)$ and the corresponding Fourier Components are shown. We use

the real and imaginary Fourier Components to form a feature vector as $v(x, y)$. These motion patterns (represented by Fourier Component) can be extracted and modeled using the Fuzzy C Means (*FCM*) method to classify each input vector $v(x, y)$ according to c fuzzy subclasses by minimizing the following sum of square errors [97]:

$$J(K_i, n) = \sum_{j=1}^n \sum_{k=1}^c (u_{ijk})^m d^2(v_{ij}, \mu_{ik}) \quad (4.16)$$

In equation (4.16), the unknowns are u_{ijk} and μ_{ik} . The meaning of each symbol is

- i : This is referring to the i^{th} expression training, i.e. one of the 6 mentioned expressions.
- K_i : The vector set for training the i^{th} expression model.
- n : The number of $v(x, y)$ in K_i . It is equal to $N_i \times N_{\%}$.
- N_i : The number of training sequences for expression i .
- c : The number of fuzzy subclasses. Experiments will be performed to find the optimal c .
- v_{ij} : The short hand form for $v_{ij}(x, y)$.
- u_{ijk} : The short hand form for $u_{ijk}(x, y)$ which is the degree of fuzzy membership of input vector $v_{ij}(x, y)$ being in the k^{th} subclass.
- m : The fuzzy exponent ($m \in [1, \infty)$), which determines the degree of overlapping between classification subclasses. If m is close to or equal 1, *FCM* becomes a hard k means clustering. On the other hand, the boundaries of

sample cluster membership becomes increasingly blurred if m is set too high.

Currently, we set m to be 2.

- μ_{ik} : The k^{th} subclass center (or mean vector) of the i^{th} expression.
- $d^2(v_{ij}, \mu_{ik})$: The distance (Euclidian distance is used here) from v_{ij} to μ_{ik} .

In equation 4.16, u_{ijk} should satisfy the following conditions:

$$\begin{aligned} u_{ijk} &\in [0, 1] \quad \forall i, j, k \\ \sum_{k=1}^c u_{ijk} &= 1 \\ 0 &\leq \sum_{j=1}^n u_{ijk} \leq n \quad \forall i, k \end{aligned} \quad (4.17)$$

According to FCM mechanism, the minimization of equation 4.16 provides an iterative solution of u_{ijk} and μ_{ik} as [97]:

$$u_{ijk} = \frac{1}{\sum_{p=1}^c \left(\frac{d(v_{ij}, \mu_{ik})}{d(v_{ij}, \mu_{ip})} \right)^{2/(m-1)}} \quad (4.18)$$

$$\mu_{ik} = \frac{\sum_{j=1}^n u_{ijk}^m v_{ij}}{\sum_{j=1}^n u_{ijk}^m} \quad (4.19)$$

The iterative algorithm is outlined in the following:

FCM Computation

Initialization: $u_{ijk} = u_{ijk}^0 = 1/c$ (Assume equal probability initially.)

$z = 1$. (z is a counter)

Repeat the following:

Begin

Update u_{ijk} and μ_{ik} using equations (5) and (6)

If error is small, i.e. $\max_{ijk} \{ |u_{ijk}^{z+1} - u_{ijk}^z| \} < \varepsilon$, then stop iteration.

*If exceed maximum number of iterations, i.e. $z > \text{max_iteration}$,
then stop iteration.*

$z = z + 1$

End repeat

The values of max_iteration and ε are set as 100 and 10^{-5} respectively. In practice, the computation converges in about 60 iterations. Typical examples of the convergences of u_{ijk} and μ_{ik} over time are shown in Fig. 4-9 and Fig. 4-10.

4.2.2.2 Expression Model Construction

Each expression model, Expr_i , composes of two sets, the motion subclasses $\{\mu_{ik}$ with $1 \leq k \leq c\}$ and the membership function matrix $\{u_{ik}(x, y)$ with $1 \leq k \leq c\}$ as

$$\text{Expr}_i = \{ \mu_{ik}, u_{ik}(x, y), \text{ with } 1 \leq k \leq c \} \quad (4.20)$$

In total, there are 6 expression types as listed in Table 4-1. Each entry of the matrix $u_{ik}(x, y)$ specifies the fuzzy membership value of subclass k for each expression. $u_{ik}(x, y)$ is related to the generated $u_{ijk}(x, y)$ from previous section. Since $u_{ijk}(x, y)$ is computed from the top $N\%$ pixels (discussed in Section 5.2.2.2 and 5.2.3.2) of significant motion from each sequence, only $N\%$ of the facial area from each sequence has valid member function values while the other areas get a value of zero.

The average values of $u_{ijk}(x, y)$ over N_i sequences will then give $u_{ik}(x, y)$ as

$$u_{ik}(x, y) = \frac{1}{N_i} \sum_{p=x, q=y} u_{ijk}(p, q) \tag{4.21}$$

where p and q are determined by j (see equation (4.16)). Fig. 4-11 shows examples of $u_{ik}(x, y)$ with $1 \leq i \leq 6$ and $1 \leq k \leq 4$ where 6 and 4 are the optimal parameters setting to be discussed in Section 5.2.2.2. Each image has been normalized by the maximum value of all the u_{ik} . These model membership matrices will be used in the recognition process to match against the input membership matrices to classify an input expression.

Table 4-1 Expression types with respect to i .

$Expr_i$					
i=1	i=2	i=3	i=4	i=5	i=6
Smile	Surprise	Anger	Disgust	Fear	Sad

4.2.2.3 Facial Expression Recognition

When the input is a test sequence for expression recognition or classification, the sequence will go through the “Feature Descriptor representation”, “Classifiers Processing” and the “Expression recognition” process in Fig. 4-1. The movement pattern at each pixel will be classified in the first process according to the pre-determined fuzzy C means, $\{\mu_{ik}$ with $1 \leq i \leq 6, 1 \leq k \leq c\}$. Consequently, the input fuzzy membership matrix $\{w_{ik}(x, y)$ with $1 \leq i \leq 6, 1 \leq k \leq c\}$ can be computed using equation (4.18). In the expression recognition process, we compute the dissimilarity value (DV_i) between the input $w_{ik}(x, y)$ and the models $u_{ik}(x, y)$ as

Fig. 4-9 Typical examples of the convergence of u_{ijk} .

Fig. 4-10 Typical examples of the convergence of μ_{ik} .

Fig. 4-11 Membership matrices of the six basic expressions.

$$DV_i = \sum_{x=1}^X \sum_{y=1}^Y D_i(x, y) \quad (4.22)$$

$$D_i(x, y) = \min_{(m,n) \in R(x,y)} \sum_{k=1}^c |w_{ik}(x, y) - u_{ik}(m, n)| \quad (4.23)$$

where X and Y are the total numbers of rows and columns of pixels inside the facial area. $R(x, y)$ is a small neighborhood centered at (x, y) with size of (1D Fourier Component: $col / 40 \times row / 40$) and (2D Fourier Component: $0 \times row / 40$).

It can be observed that different expressions have different extent of muscle movement. We assume that the largest $N\%$ pixels on the facial area form a set $S\%$ and measure the movement extent (using equation 4.9) of each sequence, k , as

$$ME(k) = \sum_{(x,y) \in S\%} Sum(x, y) \quad (4.24)$$

The means and variances of ME can be computed as

$$\mu_i = \frac{1}{N_i} \sum_{k=1}^{N_i} ME(k) \quad (4.25)$$

$$\sigma_i^2 = \frac{1}{N_i - 1} \sum_{k=1}^{N_i} (ME(k) - \mu_i)^2 \quad (4.26)$$

With an unknown expression sequence, the probability of it belonging to expression i can be computed as

$$P_i = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(ME - \mu_i)^2}{2\sigma_i^2}\right) \quad (4.27)$$

With P_i , we can refined DV_i as

$$DV_i' = DV_i * (1 - P_i) \quad (7.28)$$

Finally, the i^{th} expression match that produces the minimal dissimilarity, DV_i' , is

selected as the match.

4.3 Summary

In this chapter, we introduce our proposed local feature descriptors (1D or 2D Fourier Components) and the methods we employed to process the features (Wave Shape or Fuzzy C means). 1D Fourier Components utilize the frequency information transformed only from the time axis while the 2D Fourier Components are transformed from both the temporal and spatio space. 2D Fourier Components are essential when the temporal information is not enough (i.e. the frame number is few). For classification methods, Wave Shape method classifies the feature descriptor based on the amplitude of Fourier Components. FCM accept and handle both real and imaginary information of Fourier Components and give a probability degree for each subclass. The FCM convergence is a self-update process, which can adjust its own classification criterion itself by analyzing more and more data. This makes the system more automatic and more intelligent. Moreover, playing facial expression is a continuous process. It is hard to give quite clear and hard boundary of each expression. The blur classification given by FCM may be more close to the natural and real situation than other clear and hard classifier.

Till now, our work has explored the alternatives in building up a local descriptor from 1D or 2D Spatio-Temporal information. We will further experiment with the descriptor to search for the optimal parameter settings to build an efficient and effective system for expression modeling and recognition in next chapter.

Chapter 5 Experimental Results

To evaluate the performance of our proposed approaches, both the facial component localization and facial expression recognition systems have been built to compute recognition rates. The Cohn-Kanade facial expression database was used as the sources of the input sequences. It consists of 96 subjects, 391 expression sequences. In this database, each subject plays one or more than one of the basic six expressions, but not necessary all the six expressions. In addition, some sequences have less than the required 11 frames that exhibit movements. Hence, we cannot use all 96 subjects in our experiment. A summary of the numbers of subjects and expression sequences that were used in our experiment is tabulated in Table 5-1. In later part of this chapter, we also employ the MMI face database for cross database validation.

Table 5-1 Numbers of facial expression sequences of the Cohn-Kanade database.

Number of subjects	Smile	Surprise	Anger	Disgust	Fear	Sad
95	93	88	25	33	54	72

In later sections, first the experimental results of facial component localization is introduced, and then the results of our FER systems are displayed, which are followed

by some discussion and analysis.

5.1 Facial Component Localization

The subjects of the Cohn-Kanade database were divided into 4 subsets evenly. The proposed method has been trained and tested using a leave one subset out cross validation. We run the training and testing procedure 4 times. Each time one of the 4 subsets was used as a test set and the other 3 subsets were used as a training set.

Our facial component localization system is built on the face detection results. We have adopted the OpenCV's face detector, as discussed in section 3.2.1. The face detection rate is 100% on the Cohn-Kanade database. Such good result was also reported in [80]. The reason for this high detection rate is that all the face images were captured in a frontal view under very good illumination condition.

To evaluate the performance of the method, each of the automatically located facial components was compared to the manually located results. Assume the distance between two eye centers on the face is D , while the difference between the automatically detected and manually located results is d . If $d/D < 0.05$, we treat this localization as *success*; otherwise, the localization is regarded as *fail*. The experimental results are displayed in Table 5-2. The features are ordered by: left eye inner corner, right eye inner corner, left eye outer corner, right eye outer corner, left eyebrow corner, right eyebrow corner, left mouth corner, right mouth corner, and nose center. The average testing time, including face detection and facial component

localization, is around 50.3 ms/frame on a PC with an Intel Core2 (3.0G Hz) processor. The main reason for the mis-localization is that on the apex state of expression, the facial muscle is dramatically distorted. In such cases, our system sometimes detects facial components at incorrect position.

Table 5-2 Experimental results for facial components localization.

Feature Name	1	2	3	4	5	6	7	8	9
Frame #	6832								
Accurate #	6741	6788	6170	6257	6087	6160	6363	6343	6778
Rate (%)	98.7	99.4	90.3	91.6	89.1	90.2	93.1	92.8	99.2
Overall rate	93.8%								

Our result reported in Table 5-2 is slightly higher than Vukadinovic’s system [80], which is 93%. Both are using the Cohn-Kanade database. However, it is hard to compare our results to others work using different database fairly. Even both are using the same database for cross validation, the divisions of training/testing data are different. More importantly, the manually localized positions are also different for different users on the same face image. Vukadinovic’s system localizes 20 facial points on 300 samples while ours localizes 9 points on 6832 frames. All these differences make a fair comparison difficult.

Furthermore, the Cohn-Kanade database is not a good testing database for face detection and facial component localization. To work in a real scenario, the facial component location system needs to be able to tolerate some variations, i.e., head rotation, illumination condition changes, partially occluded face, some external objects on the face (glasses), etc. However, Cohn-Kanade database does not offer such kind of face images.

To test our facial component localization system in a complex situation, we collected 210 images from 21 subjects with different poses. These images are captured with a FLEA camera and include all the variation listed above. This time the subject-independent test gave 95% accuracy with the criterion $d/D < 0.05$ (described previously). Some examples of testing images are displayed in Appendix I. More video results for our system can be found at Youtube website:

<http://www.youtube.com/watch?v=RyallVopm9c>

<http://www.youtube.com/watch?v=voNIZzGB7ww>

Although there is no real scenario face data in the Cohn-Kanade database, we still think our results are promising. There are two reasons:

1. Our test is based on subject independent cross validation. One subject cannot appear both in the training and testing data. This is much closer to real scenario compared with subject independent test.
2. Our system localizes the facial components on all the expression states, including the non-expression state and maximum facial muscle distortion states. It is much harder to localize facial components when the facial expression is on the apex states (i.e., mouth open, eyebrow raised. As a result, in Table 5-2, features 1, 2, and 9 get the highest rate since nose/eye inner corners are relatively less variable compared with other components). Regardless of this, our algorithm still achieved 93.8% accuracy.

5.2 Facial Expression Recognition

Experiments on facial expression recognition can be conducted in two different manners: subject dependent (part of the data for each subject is used as training data while the other part is used as test) and subject independent (data of the test subject is not used as training data). Generally, subject independent test is closer to the real life scenario and happens more frequently in real applications compared with the dependent test, since it is not always possible to collect all the facial expression samples from each user, especially when the size of users is extremely large. Furthermore, the dependent test system usually achieves better performance than the independent test system. In [11], the authors reported recognition rates of 83.31% and 65.11% for the dependent and independent test respectively. Similarly, in [88], the authors tested their system on two different databases. The best dependent test results were 98.4% and 81.3%, while the best independent test results were 77.1% and 79.2%.

To perform the subject dependent test, it is necessary that the testing database has at least 2 similar expression sequences for each subject (one for training and one for testing). However, the Cohn-Kanade facial expression database has only one image sequence of each expression for each subject. Hence, we concentrate mainly on independent test.

As discussed in chapter 4, our system can be divided into two main parts: feature representation (1D and 2D Fourier Components) and classifier processing (WaveShape classifier and FCM classifier). In this chapter, three systems are tested to evaluate the proposed algorithm performance. There are: 1D_FC_WaveShape system,

1D_FC_FCM system and 2D_FC_FCM system.

In section 5.2.1, section 5.2.2 and section 5.2.3, the experiments have been carried out with the Cohn-Kanade database, using the leave one subject out cross validation. The cross database validation and bidirectional test have been shown in section 5.2.4 and section 5.2.5 respectively.

5.2.1 1D_FC_WaveShape System Results

5.2.1.1 Expression Recognition

Excluding the background, each facial area has a resolution of 200×220 . The frame numbers are different for different sequences in the Cohn-Kanade database. To include as many sequences as possible in our experiments, we accepted sequences that have at least eleven frames of expression activity. Sequences with lesser number of frames could have too little information to yield meaningful interpretation, and hence they will be excluded. If a sequence has more than eleven frames of expression activity, some frames with lesser activity will be eliminated by the procedure discussed in Section 3.4.

Our experiment was conducted in a leave one subject out cross validation manner, which is similar with [30] and [83]. We used the data of all but one subject as training data, and tested on the sequences of the subject that was left out. Each input sequence was compared with model for each of the six basic expressions. The recognition results which are not so desirable are displayed in Table 5-3.

Table 5-3 Experimental results for Cohn-Kanade database.

Input Label	Smile	Surprise	Anger	Disgust	Fear	Sadness
Smile	88	0	2	9	2	4
Surprise	0	54	0	0	0	2
Anger	0	4	22	0	2	3
Disgust	1	0	0	14	2	2
Fear	1	9	1	1	45	3
Sadness	3	21	0	9	3	58
Accuracy	94.6%	61.4%	92.0%	42.4%	83.3%	79.2%
Average	77.0%					

5.2.1.2 Comparison

Table 5-4 gives a summary of experiment results of both subject dependent and independent test from some of existing systems. It can be observed that subject depend test is always outperform the subject independent test. From Table 5-4, the experimental results tested on the Cohn-Kanade database are not as good as some others. It is expected by applying the fuzzy Spatio-Temporal approach our system can identify more complex facial expression and improve the recognition rates. More detailed results comparison can be found in Table 5-21 and Table 5-22.

Table 5-4 Subject dependent and independent system performance.

Reference	Results	
	Dependent	Independent
<i>I.Cohen et. al</i> [11]	83.3%	65.1%
<i>Ma & Khorasani</i> [110]	100.0%	93.8%
<i>W. Zheng et. al.</i> [88]	98.4%%	77.1%
Wave Shape		77.7%

Our proposed method is similar with Valstar’s work in [86]. Our method employs

Fourier Transform to represent spatio-temporal changes occurring in an image sequence. In the spatio-frequency domain, 2D images of temporal templates are constructed, which show where and which pattern motion has occurred in the image sequence. The motion patterns have been classified into 3 subclasses. Valstar used motion history images to record when and where the facial expression motion has occurred. By observing the problem of motion self-occlusion due to overwriting, Valstar proposed to record the motion history at multiple time intervals and to construct Multilevel Motion History Image (MMHI). The kNN and SNoW algorithms are employed to calculate the distance between the training samples and the test sample. Since Multilevel Motion History Image records multiple motion history, the previous motion history information will not be overwritten by later one. Compared with MMHI, our feature descriptors using Fourier Component is more focus on the motion pattern on each fixed position. In the duration of expression change, the motion may appear in “one peak”, “two peaks” and “mixture of peaks” (Fig. 4-8). We are not interested in “when” the motion has occurred since different people play the same expression in different speed. The duration time may also be different. Our system concentrates on the similar pattern that the same expression appears.

5.2.2 1D_FC_FCM System Results

5.2.2.1 Expression Recognition

A dissimilarity value was computed to measure the dissimilarity degree for each

comparison. Finally, the minimal score was selected and the corresponding expression type indicated the expression of the input data. In this test, the subclass number c was set to be 4 and 20% pixels of each sequence were selected for classification since these were the optimal settings found in the next section. The recognition results have been tabulated in Table 5-5. The proposed system has been run according to the following parameters listed in Table 5-6. From Table 5-6, it can be observed that the expression 'Smile' has achieved the best recognition rates. Michel et. al. in [82] also drew the same conclusion. The best rate is the expression having more uniform pattern among different subjects. We observed that when people smile, their mouths always moves more dramatically than when they perform others expressions. We think this makes the 'smile' expression more distinctive. For 'Disgust' and 'Sadness', the facial muscle movements can be similar to other expressions. Even we get confused and have difficulty to identify these two expressions for some sequences. This could be the reason to explain the worse performance of these two expressions. In addition, another source that contributed additional errors could be due to the fact that different subjects had played the same expression in different degrees. This unpredicted variation has made the identification more difficult and complicate. Fig. 5-1 shows several examples of 'Smile' expression with different degrees. All these four images correspond to the state with maximum muscle distortion.

Table 5-5 The best recognition results of 1D FST.

Input Label	Smile	Surprise	Anger	Disgust	Fear	Sadness
Smile	89	1	1	2	1	7
Surprise	0	79	0	1	0	3
Anger	0	3	23	0	3	3
Disgust	1	0	0	28	0	1
Fear	1	2	1	1	49	2
Sadness	2	3	0	1	1	56
Accuracy	95.7%	89.8%	92.0%	84.8%	90.7%	77.8%
Average	88.8%

Dramatically

Slightly

Fig. 5-1 Different degrees of mouth movements for the ‘Smile’ expression.

Table 5-6 Summary of the system parameters (# stands for the short form of number).

Symbol	Description	Appeared in Equation	Value
M	# of frames in each sequence	1	11
N_d	# of components used for classification	2	4
m	fuzzy exponent	3, 5, 6	2
c	# of classification subclasses	3, 4, 5	4
col	# of pixels in each column	9	220
row	# of pixels in each row	9	200

5.2.2.2 Empirical Optimal Setting

In previous section, we have discussed the recognition results with respect to the

optimal setting of c as 4 and $N_{\%}$ being the number of the top 20% pixels that showed the highest expression activity. We have empirically varied these two parameters to search for the optimal setting for the system. The recognition rates with respect to changes of these 2 parameters are tabulated and shown in Table 5-7 and Fig. 5-2. The optimal recognition rate is found to be 88.8% when c equals 4 and $N_{\%}$ equals 20%.

Table 5-7 Recognition rates with respect to c and $N_{\%}$.

Recognition Rates		Selected Pixel Percentage ($N_{\%}$)					
		5%	10%	20%	30%	40%	50%
Subclass Number (c)	2	81.1%	83.5%	87.9%	84.0%	83.2%	84.1%
	3	82.7%	84.6%	87.9%	82.7%	83.8%	84.6%
	4	82.4%	86.0%	88.8%	85.5%	85.7%	86.6%
	5	82.2%	84.6%	86.6%	83.5%	85.2%	86.0%
	6	81.1%	85.2%	85.8%	84.6%	84.4%	85.2%

Fig. 5-2 Recognition rates with respect to c and $N_{\%}$.

5.2.2.3 The Effects of Resolution and Number of Frames

The proposed system can be affected by two additional parameters, i.e. the image resolution and the number of frames per sequence. One can speculate that the performance should improve if higher resolution is used and more frames per expression sequence can be supplied since this is equivalent to supplying more and better information to the proposed system. On the other hand, one can also speculate whether it is possible to use lower resolution and less number of frames per sequence to simplify the computation process and at the same time maintain similar recognition performance without much degradation. We have performed experiments to measure the recognition rate with respect to these two parameters. The results are displayed in Fig. 5-3 and Fig. 5-4. The parameters employed in the experiment are listed in Table 5-8 and Table 5-9. Fig. 5-3 records different recognition rates according to different resolutions on the X-axis. The performance degraded gradually, and the system was able to give a recognition rate close to 0.58 even when the width was only 33 pixels wide as shown in Fig. 5-5. Fig. 5-4 displays the results with respect to different numbers of frames per sequence. The performance degraded slowly, and the system was able to give a recognition rate close to 0.6 even when the number of frames was cut down to 4. These experiments show that a working system can be built with lower resolution and smaller number of frames per sequence.

Table 5-8 Parameter setting for different resolution.

		Parameters Setting	
		$N_{\%}$	c
Resolution	200*220	20%	4
	100*110	20%	4
	67*73	20%	3
	50*55	30%	3
	40*44	30%	3
	33*36	30%	3

Table 5-9 Parameter Setting for different number of frames.

		Parameters Setting	
		$N_{\%}$	c
Number of frames	11	20%	4
	10	20%	4
	9	20%	4
	8	20%	4
	7	20%	3
	6	30%	3
	5	30%	3
	4	30%	3

Fig. 5-3 Recognition rate with respect to different resolutions.

Fig. 5-4 Recognition rates with respect to different numbers of frames.

Fig. 5-5 Examples of input images with different resolutions.

5.2.3 2D_FC_FCM System Results

5.2.3.1 Expression Recognition

The proposed system is run according to the following parameters settings in Table 5-10. The optimal setting has been found to have $M=4$, $N_p=40$, $c=16$ and $N_{\%}$ being $X \times Y \times 10\%$. Details of the findings and experiments will be described in the following sub-sections.

Our experiment was conducted in the following manner. We used the data of all but one subject as training data, and tested on the sequences of the subject that was left out. Each input sequence was compared with model for each of the six basic expressions. A dissimilarity score was computed to measure the dissimilarity degree for each comparison. Finally, the minimal score was selected and the corresponding expression type suggested the expression of the input data. In the experiment, each row was divided into 40 pieces, the subclass number c was set to be 16 and 10% pixels of each sequence were selected for classification since these were the optimal settings found (to be described in later section). The recognition results have been tabulated in Table 5-11 with an average recognition of 91.8%, which is superior as compared to other (to be discussed later). From Table 5-11, we found that the best result of 2D-FST is better than that of 1D-FST. It can also be observed that the expression ‘Smile’ achieves the best recognition rates (‘Anger’ also achieves 100% accuracy but this type of expression has much fewer sequences than ‘Smile’). This phenomenon is similar with the 1D_FC_FCM recognition results.

Table 5-10 Summary of the system parameters (# stands for the short form of number).

Symbol	Description	Appeared in Eq.	Values Tested
M	# of frames in each sequence	1, 2	2, 3, 4, ..., 11
N_p	# of pieces dividing a row	10	10, 20, 30, 40, 50
N_s	# of Fourier Components used	3	Keep 90% energy
m	fuzzy exponent	4, 6, 7	2
c	# of classification subclasses	4, 5, 6	2, 3, 4, ..., 20
$N_{\%}$	Percentage of facial area used	3	3%~40% of $X \times Y$
X	# of pixels in each column	1	220
Y	# of pixels in each row	1, 10	200
D_r	Radius of the neighborhood	11	10

Table 5-11 The best recognition results of 2D FST.

Input Label	Smile	Surprise	Anger	Disgust	Fear	Sadness
Smile	93	1	0	2	4	5
Surprise	0	77	0	0	2	4
Anger	0	0	25	0	0	1
Disgust	0	5	0	31	1	0
Fear	0	3	0	0	47	0
Sadness	0	2	0	0	0	62
Accuracy	100%	87.5%	100%	93.9%	87.0%	86.1%
Average	91.8%					

5.2.3.2 Empirical Optimal Setting

In previous section, we have generated the recognition results with respect to the optimal setting of c as 16, M as 4, N_p as 40 and $N_{\%}$ being the number of the top 10% pixels that show the highest expression activity. We have empirically varied all these 4 parameters to search for the optimal system setting. Table 5-12, Table 5-13, Table 5-14, Table 5-15, Table 5-16 and Table 5-17 record the findings with respect to

different c , N_p , and $N\%$ when $M=4$. We test the effect of M later in Section 5.2.3.3. The highest rates in each table have been highlighted.

Table 5-12 Recognition rates with respect to c and $N\%$ when $N_p=10$.

Recognition Rate		Selected Fourier Pattern Percentage ($N\%$)					
		3%	5%	10%	20%	30%	40%
Subclass Number (c)	5	65.8%	66.2%	68.1%	68.1%	66.7%	66.2%
	10	68.1%	69.1%	72.5%	71.1%	69.4%	67.7%
	12	69.1%	74.1%	78.1%	75.3%	76.3%	75.3%
	15	70.2%	75.6%	78.1%	75.6%	75.9%	73.5%
	16	70.2%	74.6%	78.1%	76.6%	75.6%	73.5%
	17	70.2%	74.1%	76.6%	76.6%	75.3%	73.2%
	20	69.1%	71.1%	75.6%	76.3%	74.1%	71.1%
	25	68.1%	70.2%	74.1%	74.1%	72.5%	70.2%
	30	79.2%	68.7%	70.2%	68.1%	67.1%	66.2%

Table 5-13 Recognition rates with respect to c and $N\%$ when $N_p=20$.

Recognition Rate		Selected Fourier Pattern Percentage ($N\%$)					
		3%	5%	10%	20%	30%	40%
Subclass Number (c)	5	72.8%	74.2%	76.1%	75.1%	74.2%	73.1%
	10	74.2%	77.1%	80.5%	79.1%	76.4%	75.7%
	12	77.1%	82.1%	84.6%	82.3%	81.3%	79.3%
	15	78.2%	83.9%	86.3%	82.6%	81.9%	80.5%
	16	79.3%	83.6%	86.3%	84.6%	83.6%	81.5%
	17	78.2%	81.3%	85.6%	83.6%	82.6%	81.2%
	20	77.1%	80.9%	83.3%	83.3%	82.3%	79.3%
	25	78.2%	80.1%	82.2%	81.2%	80.1%	78.7%
	30	74.2%	78.2%	79.3%	79.3%	78.8%	75.1%

Table 5-14 Recognition rates with respect to c and $N\%$ when $N_p=30$.

Recognition Rate		Selected Fourier Pattern Percentage ($N\%$)					
		3%	5%	10%	20%	30%	40%
Subclass Number (c)	5	77.8%	79.2%	81.1%	80.1%	79.2%	78.1%
	10	79.2%	81.1%	85.5%	84.1%	81.4%	80.7%
	12	81.1%	86.1%	88.6%	87.3%	86.3%	84.3%
	15	82.2%	88.9%	89.3%	87.6%	86.9%	85.5%
	16	83.3%	87.6%	90.4%	88.6%	87.6%	85.5%
	17	82.2%	86.3%	88.6%	88.6%	86.6%	85.2%
	20	81.1%	85.9%	87.3%	88.3%	86.3%	83.3%
	25	80.1%	82.2%	86.3%	85.2%	84.1%	82.7%
	30	79.2%	80.7%	82.2%	81.1%	80.1%	79.2%

Table 5-15 Recognition rates with respect to c and N% when Np=40.

Recognition Rate		Selected Fourier Pattern Percentage (N%)					
		3%	5%	10%	20%	30%	40%
Subclass Number (c)	5	78.7%	79.8%	81.4%	81.1%	80.1%	79.2%
	10	80.6%	84.4%	85.8%	84.1%	83.4%	81.1%
	12	84.1%	87.1%	89.9%	87.3%	86.9%	84.3%
	15	85.5%	88.6%	90.4%	88.1%	86.9%	85.5%
	16	85.9%	87.6%	91.8%	88.6%	87.6%	86.1%
	17	84.1%	86.9%	90.7%	89.9%	87.6%	85.2%
	20	81.1%	86.9%	89.3%	88.3%	87.3%	84.1%
	25	80.1%	84.1%	86.3%	86.1%	85.1%	83.3%
	30	79.2%	80.7%	82.2%	81.4%	81.1%	80.1%

Table 5-16 Recognition rates with respect to c and N% when Np=50.

Recognition Rate		Selected Fourier Pattern Percentage (N%)					
		3%	5%	10%	20%	30%	40%
Subclass Number (c)	5	76.7%	77.8%	80.1%	80.1%	79.2%	78.1%
	10	77.8%	84.1%	85.5%	84.1%	81.4%	80.7%
	12	82.2%	86.1%	89.9%	87.6%	85.9%	84.3%
	15	82.2%	86.3%	88.6%	88.9%	86.6%	84.3%
	16	83.6%	86.6%	90.1%	89.9%	87.6%	85.5%
	17	82.2%	86.6%	89.9%	89.9%	86.6%	85.2%
	20	81.1%	86.3%	89.3%	87.3%	86.3%	83.3%
	25	80.1%	82.7%	86.3%	85.2%	84.1%	82.7%
	30	79.2%	80.7%	81.1%	80.1%	80.1%	77.8%

Table 5-17 Recognition rates with respect to c and N% when Np=100.

Recognition Rate		Selected Fourier Pattern Percentage (N%)					
		3%	5%	10%	20%	30%	40%
Subclass Number (c)	5	60.3%	66.0%	66.9%	66.6%	66.3%	64.9%
	10	64.9%	68.1%	70.1%	69.0%	68.2%	67.1%
	12	65.8%	68.2%	71.1%	69.9%	69.3%	68.8%
	15	66.0%	68.5%	72.3%	71.1%	70.1%	70.1%
	16	67.1%	68.5%	73.2%	72.3%	71.8%	71.5%
	17	70.1%	74.3%	81.1%	80.0%	78.9%	76.7%
	20	78.4%	80.6%	84.9%	83.3%	83.0%	80.0%
	25	74.3%	76.2%	79.7%	78.4%	76.7%	74.3%
	30	72.3%	74.3%	76.7%	74.0%	71.1%	69.6%

As discussed in previous chapter, the right amount of information is needed for the expression descriptor to perform well. Lack of information or inclusion of redundant or irrelevant information will deteriorate the system performance. Moreover, more information takes longer time to compute the final results. From above tables, different parameters have been experimented and the optimal settings can be selected. It can be observed that the best results were achieved when 10% of the Fourier patterns were used for testing. This can be explained as only 10% of the area moves dramatically when a person plays expressions, and this area captures the main muscle movement. Furthermore, it is observed that when N_p becomes larger (the size of the descriptor becomes smaller), the performance increases. Though larger size can capture more relevant information and relationship, the pattern that can be stored in such a size can increase exponentially. This makes training very difficult if limited number of training examples is available. Hence the performance increases when the size decreases. On the other hand, the performance drops when the size is too small ($N_p = 100$) to capture enough relevant information for recognition. This is what we can observe (or explain) from the tables with respect to the setting of N_p .

5.2.3.3 The Effects of Number of Frames

The performance of the proposed system may be affected by the number of frames per sequences. If more frames are available for calculation, one can speculate that the performance should improve since more information is used. On

the other hand, if many frames (larger than a threshold) are used, it might include redundant information, which can slow down the system. It is of interest to search for this threshold such that we can use less number of frames per sequence to simplify the computation process and at the same time maintain similar recognition performance without much degradation. We have performed experiments to measure the recognition rate with respect to the employed number of frames as shown in Fig. 5-6. The number of frames started at 11 and got reduced one by one until 2. For each number of frame, we adjusted other parameters to optimize the recognition rates and chose the best results. It was observed that 1D and 2D Fourier Transforms gave different recognition rates when the number was reduced. The recognition rates for the 1D Fourier Transform dropped gradually from 11 frames to 2 frames. For the 2D Fourier Transform, the rates remained stable with the highest rate at 4 frames. The 2D results are much better since the 2D matrix contain more relevant geometric information. In addition, when the frame number reduces, less information is available in 1D matrix but this is not so severe for the 2D matrix which represents the expression information within a short line. This facilitates development of practical expression recognition systems.

Table 5-18 shows the optimal settings of parameters and timing information at each number of frames for the 2D Fourier Transform. The listed average time includes the following 3 processes: Temporal Pattern Extraction, Expression Representation and Expression Recognition. The time for loading file and preprocessing is not counted. It is obvious that the performance is quite fast.

Table 5-18 The optimal setting of parameters with respect to the numbers of frames.

		Parameters Optimal Setting				Average recognition time/sequence
		$N_{\%}$	N_p	N_s	c	
Number of Frames (M)	11	20%	40	30	20	2.07s
	10	20%	40	26	18	1.98s
	9	20%	40	22	17	1.88s
	8	10%	40	20	16	1.77s
	7	10%	40	19	16	1.67s
	6	10%	40	18	16	1.59s
	5	10%	40	16	16	1.48s
	4	10%	40	15	16	1.31s
	3	10%	40	13	16	1.27s
	2	10%	40	8	16	1.14s

Fig. 5-6 The recognition rates of 1D and 2D Fourier Transforms with respect to the numbers of frames.

5.2.4 Cross Database Validation

In above experiments, the system has been evaluated using leave one subject out cross validation. To make the experiment results more convincing, in the section below the cross database validation is used for evaluation. The Cohn-Kanade facial expression database [43] is used for parameter training while the MMI face database [114]¹ is used for testing. Since the MMI face expression video are bidirectional (neutral-apex-neutral), we manually select the apex frame and only use half of the whole video for testing (neutral-apex). The results are listed in Table 5-19. The parameters setting is the optimal setting discussed in Section 5.2.3.2.

Table 5-19 Cross database validation results.

Input Label	Smile	Surprise	Anger	Disgust	Fear	Sadness
Smile	20	2	2	1	1	1
Surprise	2	11	1	1	1	2
Anger	0	1	15	2	2	2
Disgust	0	0	1	7	4	3
Fear	0	2	3	2	14	5
Sadness	1	1	5	3	3	16
Accuracy	87.0%	64.7%	55.6%	43.8%	56.0%	55.2%
Average	60.6%					

From Table 5-19, it can be observed that the cross database validation result drops dramatically compared with the previous result in Table 5-11. After the analyzing the error cases, we found several reasons:

- 1. Unobvious Expression: Some subjects from the MMI database play their

¹ The MMI face database administrative staff informed us that they recently transferred the database to a new server so there are some problems. Till 2008/06/26, there are only partial facial expression videos are available for downloading.

expression in an imperceptible way. For example, in video “S006-108.avi”, it is very hard to capture the expression change even for human being when the subject is playing the expression “Anger”.

2. 3D Head Pose Change: When playing expression, some subjects from the MMI database move their head pose in 3-D space (i.e. “S047-004.avi”), while most of the subjects from Cohn-Kanade database change their expression only in a 2-D plane space. In such case, the parameters trained by Cohn-Kanade database are not suitable for the 3-D head movement. Pantic also pointed out this problem in [113] and suggested to use profile view face to solve the out-of-plane movement problem. We also suggest possible future direction in Section 6.3.2.
3. Similar expression: Some subjects play different expression in a similar way. For example, the expressions “Anger” and “Disgust” played by Subject 049 from the MMI database are quite confusable. Similar reasons have also been pointed out in [9] on the Cohn-Kanade database data. Better features need to be found to represent the subtle difference between these two expressions.

5.2.5 Bidirectional Test

Above experiments only involve the one way directional (neutral-apex) facial movement. In real application scenario, bidirectional facial movement (neutral-apex-neutral) is more natural. Here we evaluate the proposed system on facial

image sequences capturing the full temporal patterns of facial expression of facial expressions (neutral-apex-neutral). Since the face sequence in the Cohn-Kanade database is only one directional (neutral-apex), in this evaluation, we duplicate the Cohn-Kanade database expression sequences and “artificially” make them bidirectional (neutral-apex-neutral). Similar with Section 5.2.4, the Cohn-Kanade database is used for training while MMI face database is used for testing. The parameters setting is the same as the above optimal setting and the results are listed in Table 5-20.

Table 5-20 Bidirectional testing results.

Input Label	Smile	Surprise	Anger	Disgust	Fear	Sadness
Smile	18	2	2	2	2	2
Surprise	2	9	1	0	2	2
Anger	1	2	15	1	1	2
Disgust	0	1	3	9	6	4
Fear	1	2	2	2	11	5
Sadness	1	1	4	2	3	15
Accuracy	78.3%	52.9%	55.6%	56.3%	44.0%	51.7%
Average	56.2%					

Comparing Table 5-20 with Table 5-19, we can see that there is no big difference between one directional evaluation and bidirectional evaluation. This is because our system use Fourier Component as the feature descriptor. Fourier Component is more sensitive on the gray level values change, as discussed in Fig. 4-8. If the one directional expression (neutral-apex) and the bidirectional expression (neutral-apex-neutral) display similar motion pattern, our system can achieve similar results with one directional and bidirectional input data. However, in this thesis, only

partial face data of the MMI database have been tested in bidirectional ways. In future, we will collect more bidirectional data to make a comprehensive experiment.

5.2.6 Comparison with Other Results

It is hard to compare different system results fairly, since different database is used. Even the systems are evaluated by the same database, the number of data selected, or the cross validation division may still be different. Some recognition results from other systems are listed in Table 5-21 and Table 5-22 respectively. Together we summarize the number of testing data and evaluation methods.

In [9], 70 unknown images of 10 persons were tested for recognition of 7 expressions, i.e. the neutral expression was also tested. A mean accuracy rate of 83.3% was reported. Ma and Khorasani [110] reported a recognition result of 93.8% on a database of 60 men, each having 5 face images with the expressions neutral, smile, anger, sadness and surprise. 40 men were used for training a constructive neural network and 20 men were used for testing. The neutral expression was not tested (recognized). In [88], W. Zheng et. al reported 77.1% accuracy rates on the JAFFE databases. G. Guo and C. Dyer [111] employed 213 images of 7 expressions from 10 Japanese women as experiment data and reported an accuracy of 91.0%. In [91], Y. Wu et. al. reported an accuracy of 83.2% by classifying 183 frames from 10 subjects into 6 expression types.

Kobayashi et. al. [33] experimented with 1 human subject performing 3

expressions in 54 sequences. They achieved a recognition rate of 98.1%. In [112], 600 frames of one human subject were used for testing and a 96.8% recognition rate was obtained for 6 standard types of expressions. The 600 frames were grouped into 60 expressions with each expression spanning about 10 frames. The work reported in [30][93] tested their system performances on the Cohn-Kanade database which is also being used in this study. In [93], the authors reported that their system could achieve an 81.8% accuracy rates on facial expression recognition. In [30], the authors reported a recognition rate of 84%.

In Table 5-21 and Table 5-22, “JAFPE”, “C-K” and “U-K” represent the Japanese Female Facial Expression (JAFPE) [124], Cohn-Kanade [43] and un-known database respectively. “SD” represents that the training and testing data are from the same database while “CD” means that the Cross-Database validation has been performed.

Table 5-21 Results of recent work using static image.

Approaches	# of subjects	Expressio ns types	# of test images	Recogniti on rates	Database	Type
<i>Abboud & Davoine</i> [9]	—	7	108	83.3%	C-K	SD
<i>Ma & Khorasani</i> [110]	60	5	80	93.8%	U-K	SD
<i>W. Zheng et. al.</i> [88]	10	6	183	77.1%	JAFPE	SD
<i>G. Guo & C. Dyer</i> [111]	10	7	213	91.0%	JAFPE	SD
<i>Y. Wu et. al.</i> [91]	10	6	183	83.2%	JAFPE	SD

Table 5-22 Comparisons of recent work using sequences of images.

Approaches	# of subjects	Expression types	# of test frames or sequences	Recognition rates	Data-base	Test-Type
<i>Kobayashi et. al.</i> [33]	1	3	54 sequences	98.1%	U-K	SD
<i>Zhang & Ji</i> [112]	1	6	600 frames	96.8%	U-K	SD
<i>Pardàs et. al.</i> [30]	90	6	276 sequences	84%	C-K	SD
<i>Cohen et. al.</i> [93]	53	6	318 sequences	81.8%	C-K	SD
1D-FC-FCM	95	6	365 sequences	88.8%	C-K	SD
2D-FC-FCM	95	6	365 sequences	91.8%	C-K	SD
2D-FC-FCM	95	6	365 sequences	60.6%	C-K	CD

It is found that high recognition rates can be achieved when the numbers of human subjects, expression types and test frames (sequences) are small. Best results are achieved in [33], [110] and [112], while the number of their testing data (or testing subject) is smaller compared with others. It can be explained that small number of testing data has small variations in the feature space. Hence, it is easier for the system to find out the unique pattern in a small data set. When the test data set grows, it brings more variations to the system and makes the expression recognition process more complex.

Form this view point, in Table 5-22, it is noticed that our system has most testing sequence (365) from a large number of subjects (95). We achieved relatively high accuracy (91.8%) when the leave one subject out cross validation has been performed. As discussed above, it is hard to compare fairly with others system. Hence here we conclude that our system has a promising and interesting result.

5.3 Summary

In this chapter, the experimental results of three proposed approaches have been reported to run on the Cohn-Kanade facial expression databases and the MMI face database. The experiments were carried out in a subject independent way with leave one subject out cross validation and cross database validation. In the experiments, we also explored the relationship between the system performance and the parameters settings. We got some interesting observations and suggested some explanation of those observations. We also gave a brief error analysis on the cases of wrong decisions made by the systems.

Chapter 6 Conclusion

In this study, a fully automatic system using Spatio-Temporal information for facial expression recognition has been proposed. Experiments have been conducted and the results have been compared with others approaches. In this chapter, we summarize the work in this thesis and the contributions this study has achieved. Direction of future work will also be discussed.

6.1 Work Summary

Automatic facial expression recognition has a lot of potential application and has become an attractive research topic in computer science. In this thesis, we aim to find one accurate and effective local feature descriptor and build an automatic system to overcome some problems that the current existing systems have. Our proposed system accepts a sequence of images as input data, which displays one expression from a neutral state with no expression to a maximum muscle distortion state. Fourier Transform is employed to transform temporal information into frequency domain. The calculated Fourier Components are classified by the schemes of Wave Shape or Fuzzy

C Means (FCM). By comparing the classification results with the model data, the unknown input expression sequence type can be computed.

In [41], Robert claimed that face detection results influence the accuracy of face analysis. The recognition step afterwards can be much simplified if the input face has been normalized in both scale and orientation. In Chapter 3, we proposed to use Gabor filter as feature descriptor to precisely locate the 9 facial features. Two eye centers, which are calculated from the four eye corners, are used for the face normalization. Other facial feature (i.e. mouth, nose) positions can complement the eye corner detection when the eyes are sometimes covered by external objects (i.e. glasses).

In [64], Li mentioned that two essential issues for general pattern recognition systems: feature selection and classifier design based on selected features. Chapter 4 discusses these two issues of our system in details. We use Fourier Components (1D or 2D) to extract frequency information to build the local feature descriptors. The descriptors are classified by their wave shapes or Fuzzy C means method. At the end, Modified Hausdorff Distance is employed to help to calculate the dissimilarity score. The minimal score indicates the corresponding expressions of the input data.

Chapter 5 shows the experiment results of the three proposed systems. The experiments search for the optimal settings and recognition results of each system, compare the results with other well-known proposed systems, and test the variation of the results with respect to the changes of input image resolution, frame number, and other parameter setting.

An in-depth study of facial expression recognition on the Spatio-Temporal domain has been carried out. By testing the proposed system and comparing with others, promising results have been observed.

6.2 Contributions

The main contribution of this work presented here is to give the study on the applicability of a set of known methods (Fourier Transform, Fuzzy C Means and Hausdorff Distance) to the FER problem. Moreover, new feature descriptors are proposed to represent the features of facial components and facial expression. In this study, we favor and search for a local descriptor for expression representation. A global pattern descriptor is not appropriate since it cannot be extended to the situation when part of the face is occluded. The search tackles the need of system robustness and effectiveness with respect to the variation of the dimension and size of input data. Useful conclusions have been made, which can be applied in similar facial expression recognition system. This work has made the following contributions:

Contribution 1: We have derived a novel feature descriptors using Gabor filter for facial feature localization. These descriptors are trained by Adaboost and the most discriminate one has been selected at each stage as one weak classifier. After combining all the weak classifiers, a strong classifier is formed, which can locate the 9 facial components quickly and accurately. Based on these localization results,

the input face image can be normalized and aligned automatically. This step can reduce the human involvement and make the system more intelligent and applicable.

Contribution 2: We have designed a novel local feature descriptor using Fourier Components for facial expression recognition. This descriptor contains both spatio information and temporal information transformed by Fourier Transform. Some other systems classify the facial expression on single static image and only use spatio information. Compared with them, the temporal information describes the dynamic motion pattern, which is more robust and less influenced by outside (i.e. light) conditional changes.

Contribution 3: We have adopted the fuzzy classifier (FCM) into FER field to classify the local feature descriptor of facial expressions. Human expression change is a natural, continuous and smooth process. It is very hard to set certain and clear boundary between each other. Unlike some algorithms which execute a sharp classification in which each feature is either assigned to a class or not, FCM can make an object belong to several classes at the same time but with different degrees between 0 and 1. These intermediate memberships give more accurate representation of facial expressions compared with hard boundaries.

Contribution 4: A detailed systematic study of the proposed system has been carried out to understand the influence of different parameters setting, i.e. frame number, sub class number and image resolution. The robustness of the system have been investigated and tested. The findings provide valuable directions and hints for

system development in future work.

Contribution 5: We have integrated the face detection, facial component localization and facial expression recognition modules to build a real time facial expression recognition system. This system can be tested in real scenario.

6.3 Future Work

Although we have already built a fully automatic system for facial expression recognition and achieved promising results, there are still some limitations that can be improved. Here we list them out and set them as our future work directions.

6.3.1 2D Face Alignment

As previously discussed, automatic face feature localization is used for the input face normalization. However, only 9 feature points are not enough for describes the details of face expression movement. The work in [117] already indicated that face alignment is very important for facial expression recognition.

Cootes et al. [115] proposed two important methods for face alignment: Active Shape Model (ASM) and Active Appearance Model (AAM). Both methods use the Point Distribution Model (PDM) to constrain the face shape by Principle Components Analysis (PCA). The difference is in the feature models. In [118], Li et al. used Adaboost to model the local appearances. A histogram classifier was employed to

calculate the landmark (feature points) displacement. Experiments show that the results have been greatly improved.

Following Li's work, we plan to do the face alignment to improve the system performance of facial expression recognition. We will use the same feature described in chapter 3 to localize 53 feature points. PCA will be used to constrain the global face shape. Currently we already finish building the face alignment system and start to evaluate its performance. A demo video displays our current motion pattern detection results: http://www.youtube.com/watch?v=qpm9-l_V61Q.

Since we can localize the feature points on the contour of facial components, i.e. eyes, eyebrows and mouth, it is easy to detect their motion pattern, such as eye/mouth open and close, eyebrow raise, etc. The motion pattern detection results will help a lot for automatic facial expression recognition in future.

The work in [120] already showed good result on modeling, tracking and recognizing facial expression by using face alignment. A special ASM model is learned, in order to avoid incorrect matching due to non-linear image variations. The authors also mentioned that 3D face model provide more information for robust expression recognition while 2D shape model ignores some detailed facial deformation.

6.3.2 3D Face Alignment

As Pantic stated in [113], most existing algorithms achieve good results only

when the user faces the camera and does not change his/her 3D head pose. When the user represent out-of-plane non-rigid movements, it is hard to detect from the front view face image sequence. Pantic proposed to use multiple cameras yielding multiple views of the face to solve this problem.

Similar with Pantic's solution, in future we plan to solve this by using 3D face alignment. The input images will be captured with a stereo camera. The 3D feature point positions are calculated by simulating human binocular vision from the stereo camera images.

Gu et al. described a 3D face alignment method only using a single image in [119]. The algorithm iteratively deforms the 3D model and adjusts the 3D pose to fit the image. This is another future direction of our research.

6.3.3 Facial Expression Recognition

Our current system shows good performance when the input sequence starts from a neutral state with no expression and ends with a maximum muscle distortion state. The performance drops when the frame number reduces and when the expression does not end at the extreme muscle distortion state. In future, the emphasis should be placed on building a practical expression recognition system that takes few input frames from each expression, assuming that these two frames are not necessary the neutral state or a state with extreme muscle distortion.

Another limitation is about the expression local feature descriptor selection.

Currently we use Fourier Components as feature descriptor. Not all the feature descriptors are used in the classification. Only those with large motion energy are selected and fed to the classifiers for further processing. The selection procedures can be improved by employing statistical learning approaches (i.e., Adaboost, SVM, etc). These methods are based on the statistical analysis and can choose the most discriminable feature automatically.

Finally, in future we will implement multi-view face detection system instead of current one to tolerate more dramatic head pose variation. Afterwards, we will test our system in real scenarios to evaluate the overall performance.

6.4 Thesis Conclusion

In [42], it was claimed that temporal information can reveals more relevant information about the underlying emotional states than a single static image. Compared with the temporal moving information, the intensity information is relatively less reliable since the latter one can be severely influenced by the illumination condition of the environment. Actually, variation of illumination condition is another obstacle in the FER field. The appearance of a face expression can be dissimilar if it is observed under different light condition. The Fourier Transform can capture and transfer the temporal information into the frequency domain which records the pattern changes (Wave Form) over time. The Fourier Transform can be seen as a sum of sinusoids of different frequencies. The first

sinusoid always represents the information of the intensity information, while others sinusoids represent the magnitude of movement at the corresponding frequency. In this view, Fourier Transform is a desirable tool that can separate the intensity and frequency information. In this thesis, only the temporal movement information is used for classification. By ignoring the first sinusoid, our system has shown to be more robust and less influenced by the variation of illumination condition.

This thesis has reviewed the development of the research of facial expression recognition. Existing approaches have been classified into several groups based on the methods and experiment data they used. Various limitations of the current systems have been investigated and analyzed. In order to overcome these limitations, a novel Spatio-Temporal approach has been proposed and a system has been constructed according to the approach to test its performance. A sequence of expression images instead of just one image is analyzed in the proposed system. Three kinds of local information descriptors are proposed in this thesis: the Wave Shape Spatio-Temporal (*WST*) descriptor, 1D Fuzzy Spatio-Temporal (*1D-FST*) descriptor and 2D Fuzzy Spatio-Temporal (*2D-FST*) descriptor. The transformed frequency information is further processed by different classifiers to generate a Spatio-Temporal model for each expression type. Unknown input expressions will be matched to the models for classification. A series of thorough experiments have been conducted and the results are very promising and encouraging by comparing with others work. By analyzing the error cases and the drawback of the system, the direction of future work has also been suggested. Current work has laid a foundation for facial expression analysis.

References

- [1] M. Pantic, L.J.M Rothkrantz, "Automatic analysis of facial expressions: the state of the art", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12): 1424-1445, 2000.
- [2] N. Tsapatsoulis, K. Karpouzis, and G. Stamou, "A fuzzy system for emotion classification based on the MPEG-4 facial definition parameter", in *European Association on Signal Processing*, 2000.
- [3] D. Kim, Z. Bien, and K. Park, "Fuzzy neural networks(FNN)-based approach for personalized facial expression recognition with novel feature selection method", in *Proc. of IEEE 12th Inter. Conf. on Fuzzy Systems*, vol.(2), pp. 908-913, 2003.
- [4] S. Jung, D. Kim, K. An and M. Chung, "Efficient rectangle feature extraction for real-time facial expression recognition based on Adaboost", in *IEEE Inter. Conf. of Intelligent Robots and Systems (IROS)*, pp. 1941-1946, 2005.
- [5] W. Liu, and Z. Wang, "Facial expression recognition based on fusion of multiple Gabor features", in *18th Inter. Conf. of Pattern Recognition*, vol.(3), pp. 536-539, 2006.
- [6] M. Zubair, A. Khanam , "A personalized facial expression recognition system using case based reasoning", in *Inter. Conf. on Emerging Technologies*, pp. 630-635, 2006.
- [7] I. Kotsia, and I. Pitas, "Facial expression recognition in image sequences using Geometric Deformation Features and Support Vector Machines", in *IEEE Trans. on Image Processing*, 16(1), pp. 172-187, 2007.
- [8] G. Xue and Z. Youwei, "Facial expression recognition based on the difference of statistical features", in *8th Inter. Conf. on Signal Processing*, vol. 3, pp. 16-20, 2006.
- [9] B. Abboud, and F. Davoine, "Bilinear factorisation for facial expression analysis and synthesis", in *IEE Proc. on Vision, Image and Signal Processing*, 152(3): 327-333, June, 2005.
- [10] A. Mehrabian, "Communication without words", *Psychology Today*, 2(4): 53-56, 1968.
- [11] I. Cohen, N. Sebe, A. Garg, L. Chen and T. Huang, "Facial expression recognition from video sequences: Temporal and static modeling", *Computer Vision and Image Understanding, Special Issue on Face Recognition*, 91:

- 160-187, 2003.
- [12] P. Ekman, and T.J. Sejnowski, "NSF report – facial expression understanding", http://face-and-emotion.com/dataface/nsfrept/exec_summary.html, 2006.
 - [13] S.B., Gokturk, J.-Y. Bouguet, C. Tomasi and B. Girod, "Model-based face tracking for view-independent facial expression recognition", in *Proc. of 5th IEEE Inter. Conf. on Automatic Face and Gesture Recognition*, pp. 272-278, 2002.
 - [14] G. F. Sudha, and P. Jeyashree, "Facial expression recognition using self organizing map", in *Inter. Conf. on Computational Intelligence and Multimedia Applications*, vol. 4, pp. 219-224, 2007.
 - [15] I. Kotsia, N. Nikolaidis, and I. Pitas, "Facial expression recognition in videos using a novel multi-class Support Vector Machines variant", in *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 585-588, 2007.
 - [16] P. Kakumanu, N. Bourbakis, "A Local-Global graph approach for facial expression recognition" in *8th IEEE Inter. Conf. on Tools with Artificial Intelligence (ACTAI)*, pp. 685-692, 2006.
 - [17] G. Jiang, X. Song, F. Zheng, P. Wang and A.M. Omer, "Facial expression recognition using thermal image", in *27th IEEE Annual Inter. Conf. on the Engineering in Medicine and Biology Society (EMBS)*, pp. 631-633, 2005.
 - [18] Q. Xu, P. Zhang, W. Pei, L. Yang and Z. He, "A facial expression recognition approach based on confusion-crossed Support Vector Machine tree", in *Inter. Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pp. 309-312, 2006.
 - [19] P.S. Aleksic, and A.K. Katsaggelos, "Automatic facial expression recognition using facial Animation parameters and multistream HMMs", *IEEE Trans. Information Forensics and Security*, 1(1): 3-11, 2006.
 - [20] F. Bourel, C. Chibelushi, and A. Low, "Recognition of facial expressions in the presence of occlusion", in *Proc. of the 12th British Machine Vision Conference*, vol. 1, pp. 213-222, 2001.
 - [21] Q. Xu, P. Zhang, W. Pei, L. Yang and Z. He, "An automatic facial expression recognition approach based on confusion-crossed Support Vector Machine tree", in *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 625-628, 2007.
 - [22] S. Liao, W. Fan, A.C.S. Chung and D. Yeung, "Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features", in *IEEE Inter. Conf. on Image Processing*, pp. 665-668, 2006.
 - [23] I. Hosseini, N. Shams, P. Amini, M. S. Sadri, M. Rahmaty and S. Rahmaty, "Facial expression recognition using wavelet-based salient points and subspace analysis methods", in *Conf. on Electrical and Computer Engineering*, pp.

1992-1995, 2006.

- [24] L. He, C. Zou, L. Zhao and D. Hu, "An enhanced LBP feature based on facial expression recognition", in *27th Inter. Conf. on Engineering in Medicine and Biology Society*, pp. 3300-3303, 2005.
- [25] I. Kotsia, I. Pitas, "Real time facial expression recognition from image sequences using support vector machines", in *IEEE Inter. Conf. on Image Processing*, vol. 2, pp. 966-969, 2005.
- [26] L. Ma, Y. Xiao, K. Khorasani and R.K. Ward, "A new facial expression recognition technique using 2D DCT and k-means algorithm", in *Inter. Conf. on Image Processing (ICIP)*, vol. 2, pp. 1269-1272, 2004.
- [27] Y. Zhan, J. Ye, D. Niu and P. Cao, "Facial expression recognition based on Gabor wavelet transformation and elastic templates matching", in *Proc. of 3rd Inter. Conf. on Image and Graphics*, pp. 254-257, 2004.
- [28] B. Fasel, and J. Luettin, "Recognition of asymmetric facial action unit activities and intensities", in *Proc. of the Inter. Conf. on Pattern Recognition*, vol. 1, 2000.
- [29] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression", in *Proc. of the 4th Inter. Joint Conf. on Pattern Recognition*, pp. 408-410, 1978.
- [30] M. Pardàs, A. Bonafonte, and J. Landabaso, "Emotion recognition based on MPEG4 facial animation parameters", in *Proc. of IEEE ICASSP*, pp. 3624-3627, May, 2002.
- [31] B. Batty, and M.J. Taylor, "Early processing of the six basic facial emotional expressions", *Cognitive Brain Research*, 17: 613-620, 2003.
- [32] Y. Yacoob, and L.S. Davis, "Recognizing human facial expressions from long image sequences using optical flow", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(6): 636-642, June, 1996.
- [33] T. Kobayashi, Y. Ogawa, K. Kato, and K. Yamamoto, "Learning system of human facial expression for a family robot", in *Proc. of 16th IEEE Inter. Conf. on Automatic Face and Gesture Recognition*, pp. 481-486, May, 2004.
- [34] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 511-518, 2001.
- [35] D. Cristinacce and T. Cootes, "Facial feature detection using Adaboost with shape constraint," in *Proc. of 14th Conf. on British Machine Vision*, pp. 231-240, 2003.
- [36] P. Wang; M. B. Green, Q. Ji and J. Wayman, "Automatic eye detection and its validation", in *IEEE Inter. Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, pp. 164- 171, 2005.

- [37] Y. Gao, and M.K.H. Leung, "Line segment Hausdorff Distance on face matching," *Pattern Recognition*, 35(2): 361-371, 2002.
- [38] L. Kwan-Ho, G. Baofeng, L. Kin-Man, and S. Wan-Chi, "Human face recognition using a spatially weighted modified Hausdorff distance", in *Proc. of 2001 Inter. Symposium on Intelligent Multimedia, Video and Speech Processing*, pp. 477-480, 2001.
- [39] J.M. Leski, "Generalized weighted conditional fuzzy clustering", *IEEE Trans. on Fuzzy Systems*, 11(6): 709-715, 2003.
- [40] H. Husain, M. Khalid, and R. Yusof, "Automatic clustering of generalized regression neural network by similarity index based fuzzy c means clustering", in *2004 IEEE Region 10 Conference*, pp. 302-305, 2004.
- [41] <http://www.facedetection.com/>
- [42] J.N. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower area of the face," *Personality and Social Psychology*, 37: 2049-2059, 1979.
- [43] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis", in *Proc. of 4th IEEE Inter. Conf. on Automatic Face and Gestures Recognition*, pp. 46-53, 2000.
- [44] Y. Liu, K. Schmidt, J. Cohn, and S. Mitra, "Facial asymmetry quantification for expression invariant human identification", *Journal of Computer Vision and Image Understanding*, 91: 138-159, 2003.
- [45] E.S. Chuang, H. Deshpande, and C.Bregler, "Facial expression space learning", in *10th Pacific Conf. on Computer Graphics and Application*, pp: 68, 2002.
- [46] H. Wang, and N. Ahuja, "Facial expression decomposition", in *IEEE Inter. Conf. on Computer Vision*, 2003.
- [47] B. Abboud, F. Davoine, and M. Dang, "Facial expression recognition and synthesis based on an appearance model", *Signal Processing and Image Communication*, 10(8): 723-740, 2004.
- [48] S. Dubuisson, F. Davoine, and M. Masson, "A solution for facial expression representation and recognition", *Signal Processing Image Communication*, 17(9): 657-673, 2002.
- [49] X. Zhou, X. Huang and Y. Wang, "Real-time facial expression recognition in the interactive game based on embedded hidden Markov model", in *Proc. of Inter. Conf. on Computer Graphics, Imaging and Visualization (CGIV)*, pp. 144-148, 2004.
- [50] Y. Wang, H. Ai, B. Wu and C. Huang, "Real time facial expression recognition with Adaboost", in *Proc. of 17th Inter. Conf. on Pattern Recognition (ICPR)*, vol. 3, pp. 926-929, 2004.

- [51] M. Matsugu, K. Mori, Y. Mitari, Y. Keneda, "Facial expression recognition combined with robust face detection in a convolutional neural network", in *Proc. of Inter. Joint Conf. on Neural Networks*, vol. 3, pp. 2243-2246, 2003.
- [52] W.C. Siu, "A nesting algorithm for very fast discrete Fourier Transform", in *Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1898-1901, 1988.
- [53] Y. Yoshida, W. Benjapolakul, and A. Iwata, "An application of Hopfield network to discrete Fourier Transform", in *Inter. Joint. Conf. on Neural Networks*, vol. 2, pp. 890, 1991.
- [54] J.G. Nash, "Computationally efficient systolic architecture for computing the fourier transform", *IEEE Trans. on Signal Processing*, 53(12): 4640-4651, 2005.
- [55] O. Ersoy, "Real discrete Fourier Transform", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 33(4): 880-882, 1985.
- [56] M.P. Dubuisson, A.K. Jain, "A modified Hausdorff distance for object matching", in *Inter. Conf. on Pattern Recognition and Image Processing*, pp. 566-568, 1994.
- [57] Z. zhijia, H. Shabai, and S. zelin, "A fast strategy for image matching using Hausdorff distance", in *IEEE Inter. Conf. on Robotics, Intelligent Systems and Signal Processing*, vol. 2, pp. 915-919, 2003.
- [58] Y. Xilin, and O.I. Camps, "Line-based recognition using a multidimensional Hausdorff distance", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(9): 901-916, 1999.
- [59] C. Jin-Liang, and W. Jung-Hua, "A new robust clustering algorithm-density-weighted fuzzy c means", in *Proc. of IEEE Conf. on Systems, Man and Cybernetics*, vol. 3, pp. 90-94, 1999.
- [60] K.K. Chintalapudi, and M. Kam, "A noise-resistant fuzzy c means algorithm for clustering", in *IEEE World Congress on Computational Intelligence*, vol. 2, pp. 1458-1463, 1998.
- [61] N.R. Pal, K. Pal, J.M. Keller, and J.C. Bezdek, "A possibilistic fuzzy c means Clustering algorithm", *IEEE Trans. on Fuzzy Systems*, 13(4): 517-530, 2005.
- [62] L. Chengjia, V.M. Becerra, and J. Deng, "Extension of fuzzy c means algorithm", in *IEEE Conf. on Cybernetics and Intelligent Systems*, vol. 1, pp. 405-409, 2004.
- [63] N.N. Kachouie, J. Alirezaie, and K. Raahemifar, "Genetically derived fuzzy c means clustering algorithm for segmentation", in *IEEE Conf. on Electrical and Computer Engineering*, vol. 2, pp. 1119-1122, 2003.
- [64] S.Z. Li, L. Zhu, Z. Zhang and H. Zhang, "Learning to detect multi-view faces in real-time", in *Proc. of IEEE 2nd Inter. Conf. on Development and Learning*, pp. 172-177, 2002.

- [65] M. Yang, D.J. Kriegman and N. Ahuja, "Detecting faces in images: a survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1), pp. 34-58, 2002.
- [66] C. Huang, H. Ai, Y. Li and S. Lao, "Learning sparse features in granular space for multi-view face detection", in *7th Inter. Conf. on Automatic Face and Gesture Recognition*, pp. 401-406, 2006.
- [67] R. Lienhart, J. Maydt, "An extended set of Haar-like features for rapid object detection", in *Inter. Conf. on Image Processing*, vol. 1, pp. 900-903, 2002.
- [68] Y. Abramson and B. Steux, "YEF* real-time object detection", in *Inter. Workshop on Automatic Learning and Real-time*, 2005.
- [69] B. Wu, H. Ai, C. Huang and S. Lao, "Fast rotation invariant multi-view face detection based on real Adaboost", in *Proc. of 6th IEEE Inter. Conf. on Automatic Face and Gesture Recognition*, pp. 79-84, 2004.
- [70] T. Mita, T. Kaneko and O. Hori, "Joint Haar-like features for face detection", in *10th IEEE Inter. Conf. on Computer Vision*, vol. 2, pp. 1619-1626, 2005.
- [71] C. Liu and H. Shum, "Kullback-Leibler boosting", in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 587-594, 2003.
- [72] R. Xiao, L. Zhu and H. Zhang, "Boosting chain Learning for object detection", in *Proc. of 9th IEEE Inter. Conf. on Computer Vision*, vol. 1, pp. 709-715, 2003.
- [73] C. Huang, H. Ai, Y. Li and S. Lao, "Vector boosting for rotation invariant multi-view face detection", in *10th IEEE Inter. Conf. on Computer Vision*, vol. 1, pp. 17-21, 2005.
- [74] P. J. Phillips, M. Hyeonjoon, S. A. Rizvi and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10), pp. 1090-1104, 2000.
- [75] Q. Ji, H. Wechsler, A. Duchowski and M. Flickner, "Special issue: eye detection and tracking", *Computer Vision and Image Understanding*, pp. 1-3, 2005.
- [76] J. R. Movellan, B. Fortenberry and I. Fasel, "A generative framework for real time object detection", *MPLab technique report, University of California San Diego*, 2003.
- [77] P. Moreno, A. Bernardino and J. Santos-Victor, "Gabor parameter selection for local feature detection", in *Iberian Conf. on Pattern Recognition and Image Analysis*, pp. 11-19, 2005.
- [78] R. S. Feris, J. Gemmell, K. Toyama and V. Krüger, "Hierarchical wavelet networks for facial feature localization", in *5th Inter. Conf. on Automatic Face and Gesture Recognition*, 2002.
- [79] F. Smeraldi and J. Bigun, "Retinal vision applied to facial features detection

- and face authentication”, in *Pattern Recognition Letters*, 23(4), pp. 463-475, 2002.
- [80] D. Vukadinovic and M. Pantic, “Fully automatic facial feature point detection using Gabor feature based boosted classifiers”, in *IEEE Inter. Conf. on Systems, Man and Cybernetics*, vol. 2, pp. 1692-1698, 2005.
 - [81] B. Fasel, J. Luetttin, “Automatic facial expression analysis: a survey”, in *Pattern Recognition*, 36: 259-275, 2003.
 - [82] P. Michel and R. E. Kaliouby, “Real time facial expression recognition in video using support vector machines”, in *Proc. of the 5th Inter. Conf. on Multimodal Interfaces*, 2003.
 - [83] 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”, in *Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*, 2003.
 - [84] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel and J. Movellan, “Machine learning methods for fully automatic recognition of facial expressions and facial actions”, in *IEEE Inter. Conf. on Systems, Man and Cybernetics*, vol. 1, pp. 592-597, 2004.
 - [85] A. Kapoor, Y. Qi, R. W. Picard, “Fully automatic upper facial action recognition”, in *IEEE Inter. Conf. on Analysis and Modeling of Faces and Gestures*, pp. 195-202, 2003.
 - [86] M. Valstar, M. Pantic and I. Patras, “Motion history for facial action detection in video”, in *IEEE Inter. Conf. on Systems, Man and Cybernetics*, vol. 1, pp. 635-640, 2004.
 - [87] M. Pantic, I. Patras and M. F. Valstar, “Learning spatiotemporal models of facial expressions”, in *Proc. of Inter. Conf. on Measuring Behaviour*, pp. 7-10, 2005.
 - [88] W. Zheng, X. Zhou, C. Zou, and C. Zhao, “Facial expression recognition using Kernel Canonical Correlation Analysis (KCCA)”, *IEEE Trans. Neural Networks*, 17(1): 233-238, 2006.
 - [89] M. Yeasin, B. Bullot and R. Sharma, “Recognition of facial expressions and measurement of levels of interest from video”, in *IEEE Transactions on Multimedia*, 8(3), pp. 500-508, 2006.
 - [90] M. Pantic and Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction”, in *Proceedings of the IEEE*, 91(9): 1370-1390, 2003.
 - [91] Y. Wu, H. Liu and H. Zha, “Modeling facial expression space for recognition”, in *IEEE Inter. Conf. on Intelligent Robots and Systems*, pp. 1968-1973, 2005.
 - [92] M. Yeasin and B. Bullot, “Comparison of linear and non-linear data projection

- techniques in recognizing universal facial expressions”, in *Proc. of IEEE Inter. Joint Conf. on Neural Networks*, vol. 5, pp. 3087-3092, 2005.
- [93] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang, “Learning bayesian network classifiers for facial expression recognition using both labeled and unlabeled data”, in *Proc. of the 2003 IEEE Computer Vision and Pattern Recognition*, vol. 1, pp. 595-601, 2003.
 - [94] D. H. Kim, S. U. Jung, K. H. An H. S. Lee and M. J. Chung, “Development of a facial expression imitation system”, in *Inter. Conf. on Intelligent Robots and Systems*, pp. 3107-3112, 2006.
 - [95] J. G. Proakis, D. G. Manolakis, “*Digital Signal Processing: Principles, Algorithms, and Applications*”, Prentice Hall, 1996.
 - [96] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, in *Proc. of the 5th Berkeley Symposium on Mathematical statistics and probability*, vol. 1, pp. 287-291, 1967.
 - [97] J. C. Bezdek, “Pattern recognition with fuzzy objective function algorithms”, *Plenum Press*, New York, 1981.
 - [98] Y. Wang, “Robust object matching using a modified version of the Hausdorff measure”, *International Journal of Image and Graphics*, 3(2): 361-373, 2002.
 - [99] L. Kwan-Ho, L. Kin-Man, and S. Wan-Chi, “A new approach using modified Hausdorff distances with eigenface for human face recognition”, in *7th Inter. Conf. on Control, Automation, Robotics and Vision*, vol. 2, pp. 980-984, 2002.
 - [100] Y. Gao, “Efficiently comparing face images using a modified Hausdorff distance”, in *IEE Proc. of Vision, Image and Signal Processing*, pp. 346-350, 2003.
 - [101] D. Cheng, S. Guangda, and L. Xinggang, “Face recognition using a modified line segment Hausdorff distance”, in *2003 Inter. Conf. on machine Learning and Cybernetics*, vol. 5, pp. 2016-3021, 2003.
 - [102] Y. Lu, C.L. Tan, W. Huang, and L. Fan, “An approach to word image matching based on weighted Hausdorff distance”, in *Inter. Conf. of Document Analysis and Recognition*, pp. 921-925, 2001.
 - [103] S.H. Kim, and R.H. Park, “An efficient algorithm for video sequence matching using the modified Hausdorff distance and the directed divergence”, *IEEE Trans. on Circuits and Systems for Video Technology*, 12(7): 592-596, 2002.
 - [104] <http://www.intel.com/technology/computing/opencv/>
 - [105] http://www.intel.com/technology/itj/2005/volume09issue02/art03_learning_vision/p04_face_detection.htm
 - [106] C. Huang, B. Wu, H. AI and S. Lao, "Omni-directional face detection based on real Adaboost", in *Inter. Conf. on Image Processing*, vol. 1, pp. 593-596, 2004.

- [107] C.R. Rao, H. Toutenburg, A. Fieger, C. Heumann, T. Nittner and S. Scheid, "Linear models: least squares and alternatives, springer series in statistics", 1999.
- [108] M. Yeasin, B. Bulot, and R. Sharma, "From facial expression to level of interest: a spatio-temporal approach", in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 922-927, 2004.
- [109] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(2): 97-115, 2001.
- [110] L. Ma, and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks", *IEEE Trans. on Systems, Man and Cybernetics*, Part B, 34(3), pp. 1588-1595, June, 2004.
- [111] G. Guo, and C. Dyer, "Learning from examples in the small sample case: face expression recognition systems", *IEEE Trans. Man and Cybernetics*, Part B, 35(3), pp. 477-488, 2005.
- [112] Y. Zhang, and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(5): 699-714, May, 2005.
- [113] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of facial actions and their temporal segments from face profile image sequences", *IEEE Trans. on Systems, Man and Cybernetics - Part B*, 36(2): 433-449, 2006.
- [114] M. Pantic, M. F. Valstar, R. Rademaker and L. Maat, "Web-based database for facial expression analysis", in *Proc. IEEE Inter. Conf. on Multimedia and Expo (ICME'05)*, 2005.
- [115] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active Shape Models - Their Training and Application", *Computer Vision and Image Understanding*, 61(1): 38-59, 1995.
- [116] Y. Ma, X. Ding, Z. Wan and N. Wang, "Robust precise eye location under probabilistic framework", in *Proc. Of the 6th IEEE Inte. Conf. of Automatic Face and Gesture Recognition (FGR'04)*, 2004.
- [117] L. Zhang, H. Ai, S. Xin, C. Huang, S. Tsukiji and S. Lao, "Robust face alignment based on local texture classifiers", in *IEEE Inter. Conf. on Image Processing (ICIP-05)*, pp. 354-357, 2005.
- [118] Y. Li and W. Ito, "Shape parameter optimization for Adaboosted active shape model", in *the 10th IEEE Inter. Conf. of Computer Vision*, pp. 251-258, 2005.
- [119] L. Gu and T. Kanade, "3D alignment of face in a single image", in *the IEEE Inter. Conf. on Computer Vision and Pattern Recognition*, pp. 1305-1312, 2006.
- [120] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial

- expression”, *Journal of Image and Vision Computing*, 24(6): 605-614, 2006.
- [121] M. Pantic and M. S. Bartlett, “Machine Analysis of Facial Expressions”, in *Face Recognition*, I-Tech Education and Publishing, pp. 377-416, 2007.
- [122] Y. L. Tian, T. Kanade, and J. F. Cohn, “Facial Expression Analysis”, in *Handbook of Face Recognition*, pp. 247-276, 2005.
- [123] K. S. Huang, and M. M. Trivedi, “Robust Real-Time Detection, Tracking, and Pose Estimation of Faces in Video Streams”, in *Proc. of IEEE Inter. Conf. on Pattern Recognition*, vol. 3, pp. 965-968, 2004.
- [124] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Proc. 3rd IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

Publication List

- [1] T. Xiang, M.K.H. Leung, Y. Chen, "Facial morphing expression recognition", in *Proc. of Inter. Conf. on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, INRIA Rocquencourt, France, pp. 1–8, 2005.
- [2] T. Xiang, M.K.H. Leung, S.Y. Cho, "Expression recognition using fuzzy spatio-temporal modeling", *Pattern Recognition*, 41(1): 204-216, 2008.

Appendix I: Examples of Facial Components Localization

In the following part, some images are displayed to illustrate our facial feature localization (FCL) results in different situation. The results are divided into two parts: successful localization and unsuccessful localization.

Successful Localization

Fig.-A 1: FCL with glasses-small face.

Fig.-A 2: FCL with glasses-large face.

Fig.-A 3: FCL with glasses-smile.

Fig.-A 4: FCL with glasses-surprise.

Fig.-A 5: FCL with glasses-anger.

Fig.-A 6: FCL with glasses-rotated head-unsymmetrical illumination.

Fig.-A 7: FCL with glasses-strong illumination.

Fig.-A 8: FCL with glasses- partially occluded face (1).

Fig.-A 9: FCL with glasses- partially occluded face (2).

Fig.-A 10: FCL with glasses-partially occluded face (3).

Fig.-A 11: FCL without glasses-eye closed.

Fig.-A 12: FCL without glasses-partially occluded face.

Fig.-A 13: FCL without glasses-rotated head.

Unsuccessful Facail Components Localization

Fig.-A 14: FCL without glasses-rotated head (1).

Fig.-A 15: FCL without glasses-rotated head (2).

Fig.-A 16: FCL with glasses-rotated head.

Fig.-A 17: FCL with glasses-glisten on the glasses.