Facial Expression Recognition: Fusion of a Human Vision System Model and a Statistical Framework

Gu Wenfei

Department of Electrical & Computer Engineering

National University of Singapore

A thesis submitted for the degree of

Doctor of Philosophy (PhD)

May 18, 2011

Abstract

Automatic facial expression recognition from still face (color and gray-level) images is acknowledged to be complex in view of significant variations in the physiognomy of faces with respect to head pose, environment illumination and person-identity. Even assuming illumination and pose invariance in face images, recognition of facial expressions from novel persons always remains an interesting and also challenging problem.

With the goal of achieving significantly improved performance in expression recognition, the proposed new algorithms, combining bio-inspired approaches and statistical approaches, involve (a) the extraction of contour-based features and their radial encoding; (b) a modification of HMAX model using local methods; and (c) a fusion of local methods with an efficient encoding of Gabor filter outputs and a combination of classifiers based on PCA and FLD. In addition, the sensitivity of existing expression recognition algorithms to facial identity and its variations is overcome by a novel composite orthonormal basis that separates expression from identity information. Finally, by way of bringing theory closer to practice, the proposed facial expression recognition algorithm has been efficiently implemented for a web-application.

Dedicated to my loving parents, who offered me unconditional love and support over the years.

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my supervisor and mentor, Professor Xiang Cheng. His wide knowledge and logical way of thinking have been of great value for me. His understanding, encouraging and personal guidance have provided a good basis for the present thesis.

I wish to express my warm and sincere thanks to Professor Y.V. Venkatesh, for his detailed and constructive comments, and important support throughout this work. His enthusiasm for research has greatly inspired me.

I shall extend my thanks to graduate students of control group, for their friendships, support and help during my stay at National University of Singapore.

Finally, my heartiest thanks go to my parents for their love, support, and encouragement over the years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans recognize facial expressions with deceptive ease because, the researchers so contend, they have brains that have evolved to function in a three-dimensional environment, and developed cognitive abilities to make sense of the visual inputs. Since the precise underlying mechanisms of human recognition of patterns are not known, it has been found to be extraordinarily difficult to build machines to do such a job. Many reasons have been adduced to account for this limitation: significant variations in the physiognomy of faces with respect to head pose, environment illumination, person-identity and others. Normal color (and gray-level) face images, while exhibiting considerable variations, contain redundant information in intensity for describing facial expressions. A face image by itself has not been successfully employed in expression recognition in spite of normalization techniques to achieve illumination, scale and pose invariance. The implication is that appropriate features are needed for facial expression classification, as, in fact, evidenced by the observed human ability to recognize expressions without a reference to facial identity [11, 63].

It has been found that facial expression information is usually correlated with identity [7] and variations in identity (which are regarded as extrapersonal) dominate over those in expression (which are regarded as intrapersonal). This brings us to an unresolved, and hence challenging, problem: How to automatically recognize expressions of a novel (i.e., a face not in the database) person? In spite of many years of research, designing a system to recognize facial expressions has

remained elusive. In the following, a brief overview of researches on facial expression recognition using both statistical and bio-inspired approaches will be provided.

1.1 Overview

The problem of facial expression recognition has been subjected mostly to statistical approaches [14], which treat an individual instance as a random vector, apply various statistical tools to extract discriminating features from training examples, and then classify the test vector using its features. Significant success has already been achieved by such a strategy, and learning machines have been developed to recognize facial expression, speech, fingerprint, DNA sequence and others.

How then do such machines compare with human brains? It is found that many aspects of learning capability of humans - the most obvious one is the human ability to learn from a few examples - cannot be captured by statistical theory. For instance, in the case of recognition of objects by a machine, the number of training examples needed runs into hundreds to ensure satisfactory performance. While this number is small compared to the dimensions of the image (usually of the order of 10^6 pixels), even a small child can learn the same task from just a few examples.

Another major difference (between machines and humans) is the ability to deal with large (statistical) variance in the appearance of objects. Humans can easily recognize facial expressions of different persons, under different lighting conditions, and in different poses; understand spoken words; and read handwritten characters - all these have turned out to be extremely difficult for machines built on statistical principles.

Therefore, two natural questions arise: What is missing in the learning machines? How can we make them "intelligent", if intelligence implies, in our case, recognition of visual patterns? A typical answer to the first question by many scientists is that the human brain computes in an entirely different way from a conventional digital computer does. The answer to the second one has been the Holy Grail of the engineering community.

It is our strong belief that a new, *bio-inspired machine paradigm*, which incorporates the essential features of a biological learning system in a statistical framework, is needed to enhance the pattern recognition ability of present-day machines to a level comparable to that of human beings.

1.2 Statistical Approaches

1.2.1 Principal Component Analysis

Principal component analysis (PCA) [58], is one of the common statistical methods used in pattern recognition. Depending on the field of application, it is also called the discrete Karhunen-Lo'eve transform (KLT), or the Hotelling transform, and has been widely used in face and facial expression recognition [41, 57, 59, 79].

Suppose that there are n d-dimensional sample images $x_1, \ldots, x_n$ belonging to C different classes with n_i samples in the class Ω_i, $i = 1, \cdots, C$. Here n is the sample size and d is the dimension of feature vectors. PCA seeks a projection matrix W that minimizes the squared error function:

$$J_{PCA}(W) = \sum_{k=1}^{n} ||x_k - y_k||^2 \tag{1.1}$$

where $y_k = W(W^T x_k)$ is obtained after projection of x_k by W, and n is the total number of samples. The solution is the eigenvector of the total scatter matrix defined as:

$$S_T = \sum_{k=1}^{n} (x_k - \mu)(x_k - \mu)^T \tag{1.2}$$

where μ is the mean of all the samples:

$$\mu = \frac{1}{n} \sum_{k=1}^{n} x_k. \tag{1.3}$$

The main properties of PCA are: approximate reconstruction, orthonormality of the basis, and decorrelated principal components. That is to say,

$$x \approx Wy \tag{1.4}$$

$$W^T W = I \tag{1.5}$$

$$YY^T = D \tag{1.6}$$

where Y is a matrix whose kth column is y_k, and D is a diagonal matrix.

Usually, the columns of W associated with significant eigenvalues, called the principal components (PCs), are regarded as important, while those components with the smallest variances are regarded as unimportant or associated with noise. By choosing m ($m < d$) important principle components, the original d-dimensional vectors are projected to m-dimensional space. The resulting low dimensional vectors preserve most information and thus can be used as feature vectors for facial expression recognition.

PCA is mathematically a minimal mean-square-error representation of a given dataset. Since no prior knowledge is employed in such a scheme, PCA can be considered as an unsupervised linear feature extraction method that is largely confined to dimension reduction.

One of the limitations of the PCA is that it may not be able to find significant differences between training samples relevant to different classes if the differences appear in the high order components. This is due to the fact that PCA maximizes not only the between-class scatter which is useful for classification, but also the within-class scatter which is redundant information. For example, if PCA is applied to a set of images with large variations of illuminations, the obtained principal components preserve illumination information in the projected feature space. As a result, the performance of PCA on facial expression recognition is unstable with large variations in illumination conditions. Another problem of PCA is that it cannot separate the differences between face identities and facial expressions which are correlated with each other in the face images. Therefore, when recognizing expressions from a novel face, the performance of PCA based facial expression recognition is significantly lower than that of recognizing expressions from known persons.

1.2.2 Fisher's Linear Discriminant Analysis

Fisher's linear discriminant (FLD) analysis , a classical technique first proposed by Fisher to deal with two-class taxonomic problems [19], enables us to extract discriminating features based on prior information about classes. Even though it has been extended to multi-class problems, as described in standard textbooks on

pattern classification [14, 21, 53], it was not as popular as the PCA for extracting discriminating features until about 15 years ago. As applied to the problem of face recognition, comparisons have been made between FLD analysis and PCA in [4, 16, 72], in which it has been demonstrated that FLD analysis outperforms PCA. FLD analysis and its variants [52, 66, 71] have also shown outstanding performance with respect to facial expression recognition.

Let the n d-dimensional feature vectors under consideration be represented by $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$. Let the number of classes be C, and the number of vectors in class Ω_i be n_i, for $i = 1, 2, \cdots, C$. The FLD analysis maximizes the following cost function:

$$J(w) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}, \tag{1.7}$$

where $\mathbf{w}$ is a d-dimensional vector; and the between-class scatter-matrix S_B and the within-class scatter-matrix S_W are defined by

$$S_B = \sum_{i=1}^{C} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T, \tag{1.8}$$

$$S_W = \sum_{i=1}^{C} \sum_{\mathbf{x} \in \Omega_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \tag{1.9}$$

and

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \Omega_i} \mathbf{x}, \quad \text{and} \quad \mathbf{m} = \frac{1}{n} \sum_{i=1}^{C} n_i \mathbf{m}_i. \tag{1.10}$$

The corresponding generalized eigenvalue problem is: solve for λ and $\mathbf{w}$ from the equation,

$$S_B \mathbf{w} = \lambda S_W \mathbf{w}. \tag{1.11}$$

Since the rank of S_B is at most $C - 1$, the number of non-zero eigenvectors $\mathbf{w}$ is at most $C - 1$. Hence the dimension of the projected feature vectors is at most $C - 1$.

In facial expression recognition, it is normally the case that the sample size n is much smaller than the feature dimension d. As a result, S_w is singular, and Equation 1.11 cannot be solved. To address this issue, an indirect but effective approach [4] is to employ PCA first to reduce the feature dimension so that S_w becomes non-singular. Subsequently, FLD analysis is invoked for classification.

On the other hand, although FLD analysis can improve the performance of facial expression recognition when the images are from known persons, the recognition accuracy of expressions from novel faces has been found to be unsatisfactory due to the correlations between identity and expression found in the features currently used for expression classification.

Against the above background of a possible dichotomy between facial identity and expression, a motivation for the proposed bio-inspired approaches is the highly sophisticated human ability to perceive facial expressions, independent of identity. Though the underlying biological mechanism for this ability has not yet been understood, it seems to be expedient to study some models of the human vision system which we consider in the next section.

1.3 Human Vision System

1.3.1 Structure of Human Vision System

The human vision system processes visual signals falling on the retina of human beings and represents the three-dimensional external environment for cognitive understanding [33]. At the beginning, the retina converts patterns of light into neuronal signals. These signals are processed in a hierarchical fashion by different parts of the brain, from the retina to the lateral geniculate nucleus, and then to the primary and secondary visual cortex of the brain, resulting in two visual pathways: the dorsal stream - dealing with motion analysis, and the ventral stream - dealing with object representation and recognition [26]. The ventral stream starts with primary visual cortex and goes through visual area V2 and V4, and to the inferior temporal (IT) cortex. These visual areas are critical to object recognition and will be introduced below.

1.3.2 Retina

Cells in the retina, called retinal ganglion cells, receive and translate light into nerve signals and begin the preprocessing of visual information. Each receptive

field [1] of retinal ganglion cells composes of a central disk and a concentric ring, responding oppositely to light. This kind of receptive field enables retinal cells to convey information about discontinuities in the distribution of light falling on the retina, which often specify the edges of object.

1.3.3 Primary Visual Cortex (V1)

Generally, receptive fields of cells in V1 are larger and have more complex stimulus requirements than those of retinal ganglion cells [34]. And these V1 cells mainly respond to stimulus which are elongated with certain orientations. Moreover, V1 keeps the spatial information of visual signals from retinal cells, which is called retinotopic representation. However, this representation is distorted in the cortical area such that the retinal fovea is disproportionately mapped in a much larger area of the primary cortex than the retinal periphery [55]. In fact, V1 cells extract low-level local features of the visual information, by highlighting the lines with different directions in the visual stimulus.

1.3.4 Visual Area V2 and V4

Visual area V2 and V4 are the next stages which further process the visual information. Functionally, receptive fields of cells in V2 have similar properties to those in V1 such that cells in V2 are also tuned to stimulus with certain orientations. On the other hand, cells in V4 respond to intermediate features, such as corners and simple geometric shapes. Cells in V4 combine the low-level local features into intermediate features according to their spatial relationships, and these intermediate features are fed in to higher-level visual areas for post-processing. This kind of hierarchical procedure enables human beings to efficiently recognize different kinds of objects in a complex environment.

[1] Generally, the receptive field of a neuron is a region of space in which the presence of a stimulus will alter the firing of that neuron.

1.3.5 Inferior Temporal Cortex (IT)

Inferior temporal cortex, one of the higher levels of the ventral stream of human vision system, is associated with representation of complex object features, such as global shapes. Cells in IT respond selectively to a specific class of objects, such as faces, hands, and animals. More specifically, researchers [76, 77, 78] discovered that cells in a certain sub-area of IT, called fusion face area (FFA), receive visual information, consisting of intermediate features from the previous visual areas, and respond mainly to faces, especially to facial identities. Later, cells in another sub-area, called superior temporal sulcus (STS) process the visual information after FFA and respond mainly to facial expressions. This infers that the facial identity information would be separated from the facial expression information such that the universal expression features, which may contribute to improving the performance of facial expression recognition, could be extracted by cells in STS.

1.4 Bio-Inspired Models Based on Human Vision System

Based on the human vision system, many biologically plausible models of human object recognition have been proposed [22, 24, 61, 83], among which the following simplified three-stage hierarchical structure of the visual cortex seems to be a dominant theme:

1. Basic units, such as simple cells in the V1 cortex, respond to stimuli with certain orientations in their receptive fields, thereby extracting low-level local features of the stimuli.

2. Intermediate units such as cells in the V2 and V4 cortex, integrate the low-level features extracted in the previous stage, and obtain more specific global features.

3. Decision-making units recognize objects based on the global features.

In the following, a few bio-inspired models that play an important role in our proposed (expression recognition) scheme will be introduced, including 1) Gabor filters, imitating the V1 cells; 2) local methods, inspired by the local feature extraction and processing scheme in human vision system; and 3) hierarchical max (HMAX) model, simulating the feed-forward structure of V1 - V4 visual areas and dealing with the simple object recognition task.

1.4.1 Gabor Filters

Gabor filter, proposed by Daugman [12] and Jones and Palmer [38], has been found to be a very successful model, imitating the spatial orientation properties of cells in the V1 cortex. When convolved with an image, Gabor filters produce outputs that are robust to minor (i) object rotation and distortion; and (ii) variations in illumination.

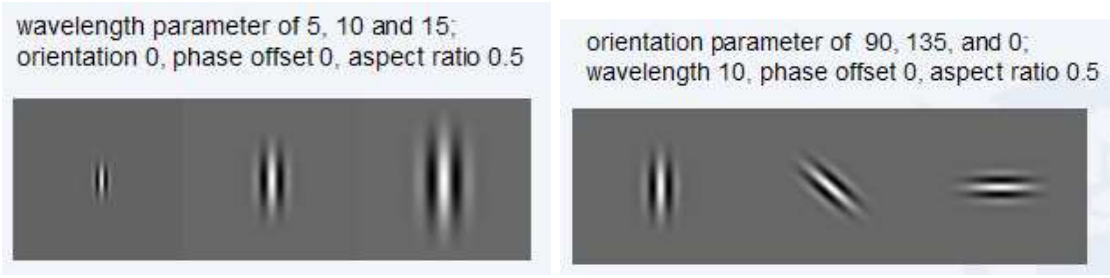
Mathematically, a set of Gabor filters can be described by the following equations:

$$g_{\lambda,\theta,\phi,\alpha,\gamma}(x,y) = exp(-\frac{x'^2+\gamma^2 y'^2}{2\alpha^2})cos(2\pi\frac{x'}{\lambda}+\phi) \tag{1.12}$$

$$x' = xcos\theta + ysin\theta, y' = -xsin\theta + ycos\theta \tag{1.13}$$

and where (x, y) refers to the pixel position in a $2D$ coordinate system, and the parameters affecting the filter outputs are: θ (orientation), γ (aspect ratio), σ (effective width), φ (phase), and λ (wavelength). These parameters can be chosen such that the filters model the tuning properties of V1 cells. Figure 1.1 (a) shows Gabor filters with different wavelength values for fixed orientation, phase offset, aspect ratio and effective width; Figure 1.1 (b) shows Gabor filters with different orientations for fixed wavelength, phase offset, aspect ratio and effective width. Fig 1.2 shows the outputs of a convolution operation on a face image with Gabor filters. It is found that Gabor filters with (i) different orientations highlight different edges; and (ii) different effective widths extract different details of information.

However, the Gabor filter outputs, when used as features for facial expression recognition, are found to contain redundant information at neighboring pixels.



(a) (b)

Figure 1.1: (a) Left: Gabor filters with different wavelength and other fixed parameters; (b) Right: Gabor filters with different orientations and other fixed parameters.

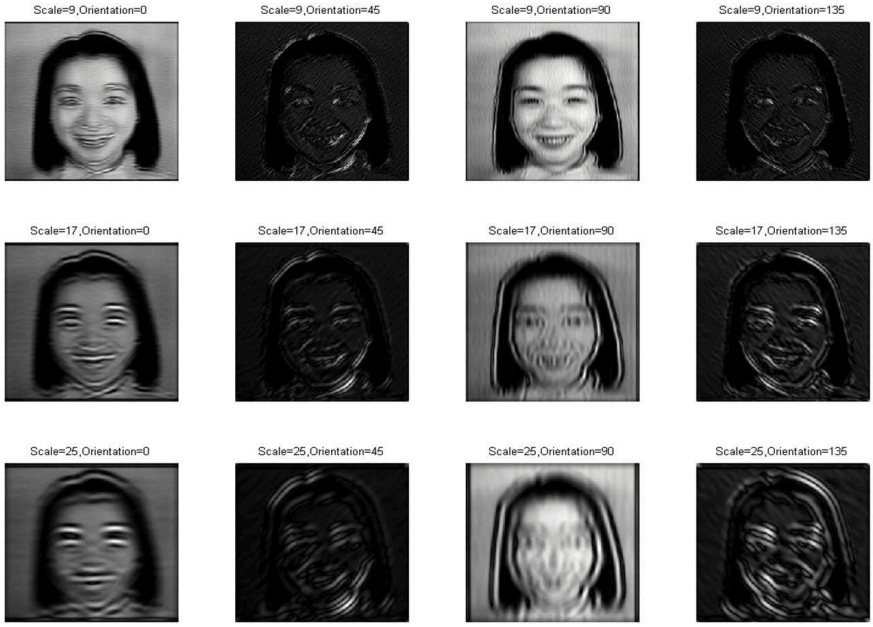


Figure 1.2: The outputs of convolving Gabor filters with a face image.

To address this issue, Gabor jets [60] have been introduced to statistically post-process the Gabor outputs to arrive at salient features. All the Gabor outputs with different parameters at one image location form a jet. There are generally two kinds of Gabor jets: selected fiducial points and uniformly downsampling. The first kind involves the choice of Gabor filter outputs at manually selected (fiducial or) interested points on the face image (such as eyebrows, eyes, nose

and mouth) [91]. In the second kind of Gabor jets, the Gabor filter outputs are uniformly downsampled by a chosen factor, and the resultant outputs are used to represent information in a facial expression [13].

The problem with the first kind of Gabor jets is that the manual selection of points for generating Gabor features makes the whole procedure non-automatic. Even though some algorithms have been proposed to automatically select feature points, the performance is still not satisfactory compared to manual interaction. Similarly, the uniformly downsampling method is limited by the choice of the downsampling factor. Too large a downsampling factor may lose critical feature points while too small a downsampling factor may not reduce the redundant information. Therefore, an efficient encoding strategy for Gabor outputs is needed to extract useful facial expression information. And this provides a motivation for our proposed scheme.

1.4.2 Local Methods

As suggested by recent physiological studies [76, 77, 78], face processing is performed by dedicated machinery in the human brain, and is believed to consist of the following:

1. Face detection and its simultaneous identification, and further processing for its expression recognition.

2. Capturing local facial information in each cell acting as a local receptive field.

3. Possible reconstruction of a face, preserving most facial information, by combining local information.

The concept of a local receptive field has led to local matching methods based on local facial features for face recognition. PCA has been applied not only to the whole face but also to the facial components, such as eyes, noses and mouths [59], resulting in a combination of eigenfaces and other eigenmodules. In [27], it is argued that local facial features are invariant to moderate changes in pose, illumination and facial expression, and, therefore, the face image should be

divided into smaller local regions for extracting local features. Even an adaptively weighted sub-pattern PCA has been proposed [73] for local regions since different human facial components may have different contributions to face recognition.

For extracting discriminating local features, the elastic bunch graph matching (EBGM) method of [84], converts a face image to a graph structure, and attaches a set of Gabor-filtered facial components to a number of nodes of the graph. New faces are recognized by comparing the similarity of both nodes and topography of the generated graphs. For a local binary pattern (LBP) based face description, see [1] in which a facial image is first divided into several local blocks, and LBP descriptors are applied to each block independently. The occurrences of the LBP codes in each block are converted into a histogram, and then combined together to build the global feature histogram. Experimental results seem to show that the LBP feature-based method is more robust against variations in pose or illumination than holistic methods. In [90], Gabor filters with five scales and eight orientations are first applied to the face image, followed by a local binary operation on the resulting 40 Gabor filtered images to obtain the local Gabor binary pattern histogram sequence (LGBPHS). New faces are recognized by comparing their LGBPHS with that of the reference faces.

Face recognition performance seems to be significantly improved when local features are employed, in comparison with that using only holistic features, as reported in [31] and [93]. Hence it is believed that local methods can produce promising results in both facial identity and expression recognition. Therefore, more experiments need to be performed in order to demonstrate the capability of local methods on facial expression recognition.

1.4.3 Hierarchical-MAX (HMAX) Model

Concerning the process of rapid object recognition in the human visual cortex, there exists the successful hierarchical MAX model (HMAX) [61] which can be briefly described as follows:

- The hierarchical visual processing consists of a series of stages that have increasing invariance to object transformations;

- As the receptive fields of the neurons increase along the visual pathway, the complexity of their preferred stimuli increases;
- Learning is probably involved at all stages and unsupervised learning may occur at the intermediate layers while supervised learning may occur at the top-most layers of the hierarchy.

1.4.3.1 Standard HMAX Model

In the standard HMAX model, there exist a number of layers of computational units. Simple S units tune to their inputs using a bell-shaped function to achieve pattern matching, while the C units perform the max operation on the S level responses. As shown in Figure 1.3, the first layer of HMAX, S1, imitating the simple cells found in the V1 area of the primate brain, consists of Gabor filters, tuned to stimuli with different orientations and scales in the different areas of the visual field. Then, the C1 units in the next layer perform max operation over the outputs of the S1 filters that have the same orientation, but different scales and positions over some neighborhoods. And in the S2 layer, composite features are obtained by combining the simple features from the C1 layer (with different orientations) into a 2 by 2 matrix of arrangements. Finally, every C2 layer unit pools the max response over all S2 units in different positions and scales, resulting in a specific feature which is used for classification. Such an architecture of multiple S and C levels enables the HMAX to increase specificity in feature detectors, and improves invariance to moderate scale and position changes. Experimental results show that HMAX model performs well when recognizing paperclip-like objects since features in HMAX were obtained by combining 4 bar orientations into 2 by 2 forms.

The HMAX architecture is supported by experimental findings on the ventral visual pathway in the primate brain, and the computational results seem to be consistent with those of physiological experiments on the primate visual system.

1.4.3.2 HMAX Model with Feature Learning

Since the intermediate features in HMAX are manually determined, the features turn out to be the same for all object classes. And since these features are ob-

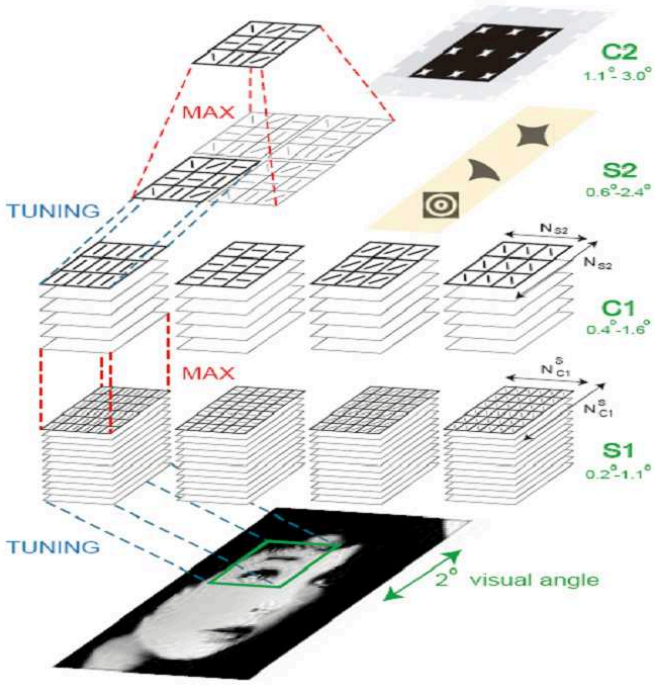


Figure 1.3: The structure of standard HMAX model [61].

tained by combining 4 bar orientations into 2 by 2 matrix forms, they may work well for paper-clip-like objects but not for face images. To address this issue, a feature learning strategy, which corresponds to selecting a set of N prototypes $\mathbf{P}_i$ (or features) for the S2 units, has been applied to the standard HMAX model to obtain class-specific features [64]. The learning is achieved by extracting a set of patches with various sizes and at random positions from training set. As shown in Figure 1.4, a patch $\mathbf{P}$ of size $n \times n$ contains $n \times n \times 4$ elements which can be extracted at the level of the C1 layer across all 4 orientations. These prototypes replace the S2 features in the standard HMAX. Then new S2 units, acting as Gaussian RBF-units, compute the similarity scores (i.e., Euclidean distance) between an input pattern X and the stored prototype $\mathbf{P}$: $f(X) = exp(-\frac{||X-P||^2}{2\sigma^2})$, with σ chosen proportional to patch size. HMAX with RBF-like feature learning has been successful in automatic object recognition, because the performance of HMAX with feature learning on rapid object recognition is similar to that of human beings. Louie [51] applied HMAX with feature learning to face detection

in cluttered background, see [51] in which a high performance has been reported.

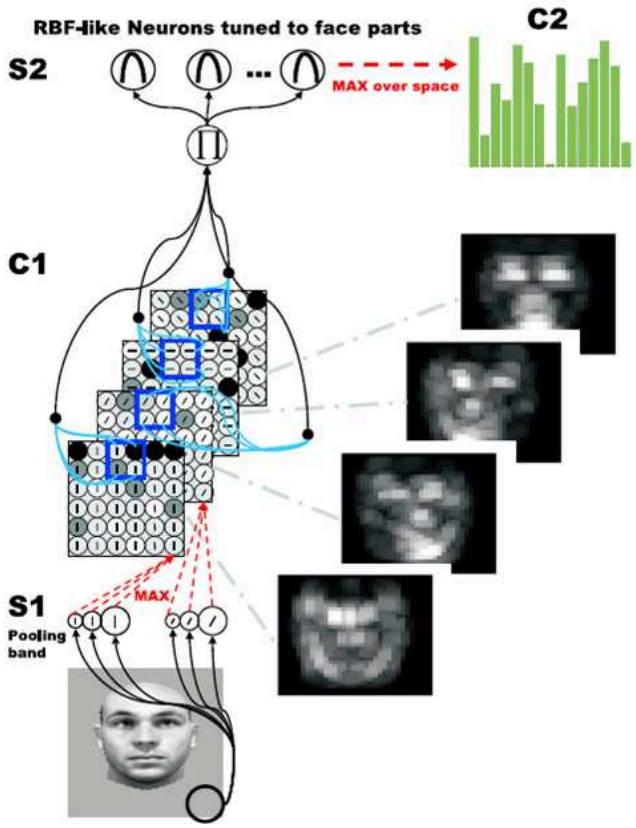


Figure 1.4: The structure of HMAX with feature learning [64].

1.4.3.3 Limitations of HMAX on Facial Expression Recognition

Even though the HMAX model with feature learning can produce strong preferences to faces against natural scenes, it cannot deal with facial expression recognition satisfactorily because HMAX cannot capture crucial properties of facial expression for the following reasons:

1. Special units to deal with face processing are missing. In standard HMAX, the final layer with C2 units, modeling the cells in the IT area, responds to a series of complex visual forms. However, according to the human vision system (e.g., cells in FFA), facial patterns are so complicated that an additional layer is needed for face processing.

2. The feature learning algorithm of HMAX generates a number of random patches which are then used as the prototypes of different objects. To achieve satisfactory performance on object classification using this kind of learning strategy, a large number of natural images are required to train the system. Although the trained system is able to respond to faces, it cannot capture detailed facial information. Therefore, HMAX can at most act as a face detector but cannot distinguish among either individual faces or different expressions.

3. Since the HMAX is trained using a set of face images with different identities and expressions, the strong responses of C2 units may correspond only to some local facial components due to the randomness of the learning strategy and the max operation of the C2 units. Therefore, the final decisions of identities and expressions may be not reliable.

1.5 Scope and Organization

The limitations of existing algorithms for facial expression recognition are summarized below to provide the background for the proposed fusion of human vision system model and statistical approaches.

1. Algorithms based on PCA and FLD analysis require large training samples to extract features (meant for discriminating expressions). But the available training samples are small in number when compared with the dimension of the training data.

2. Bio-inspired models, such as Gabor filters and HMAX, may exhibit good performance on object recognition. However, the encoding strategy involving Gabor filters is inefficient, while HMAX is not applicable to facial expression recognition.

3. Facial expression is normally correlated with identity, and variations in identity dominate over those in expression. Existing algorithms, which seem to perform well on person-dependent expression recognition, are substantially less efficient on person-independent expression recognition.

Motivated by the above, a new framework for facial expression recognition, fusing statistical approaches with bio-inspired models, is proposed in this study. More specifically, the detailed components of the proposed framework are listed in the following:

- A contour-based facial expression recognition algorithm whose performance is close to that of humans.
- Modification of the HMAX model using local methods to recognize facial expressions from novel faces.
- A composite orthonormal basis (COB) algorithm to separate the problem of recognizing expression from that of identity.
- A new facial expression recognition framework, incorporating (a) local methods, (b) efficiently encoded Gabor filters and (c) PCA and FLD analysis based classifier synthesis.
- An efficient web-application of facial expression recognition system based on the proposed framework.

As illustrated in Figure 1.5, first of all, the contour-based facial expression recognition algorithm, inspired by retinal ganglion cells, is proposed to recognize expressions using bio-plausible expression features, such as contours. Secondly, the standard HMAX model is modified by adding facial expression processors, which incorporates local methods and Gabor filters, according to the recent biological researches on FFA cells. Thirdly, the COB algorithm is proposed to imitate the cells in STS, which separate identity information from expression information, resulting in universal expression features that may lead to improved facial expression recognition performance. Finally, a new facial expression recognition framework, as well as its web-based implementation, combining improved bio-inspired models and traditional statistical approaches, is proposed for achieving elegant performance on recognizing facial expressions from novel faces.

The results of this present study may shed light on developing real-time facial expression recognition system with improved recognition accuracy when recognizing expressions from novel persons:

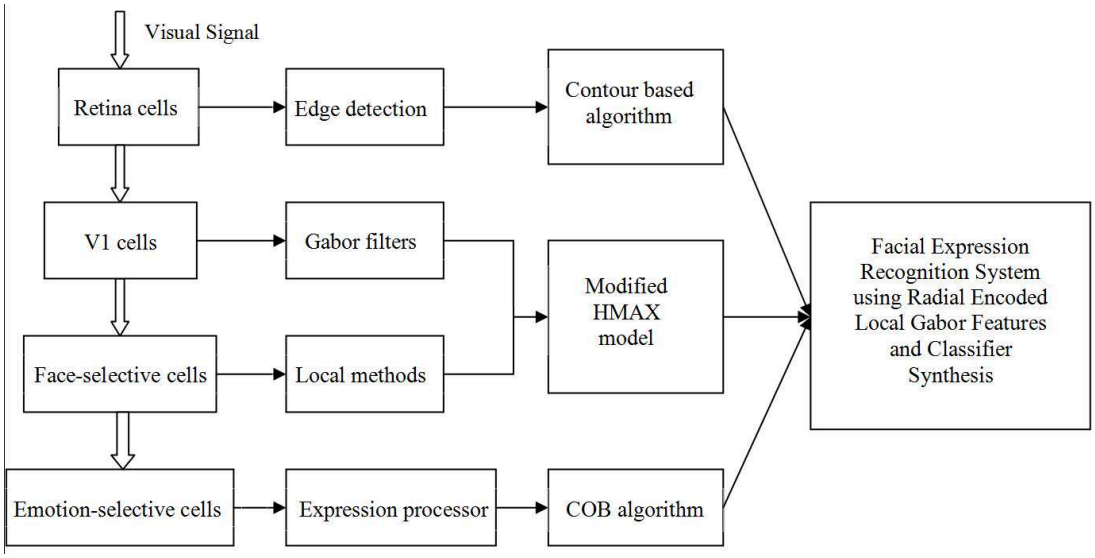


Figure 1.5: The general block-schematic of proposed algorithms simulating the human vision system.

1. the simplified 3-stage hierarchical structure of human visual cortex should be useful for designing the framework of facial expression recognition system;

2. the composite orthonormal basis should remove identity information as much as possible from face images and the resulting expression features should be discriminating for facial expression recognition;

3. the radial grid encoding strategy based on retinotopic mapping should be able to efficiently downsample the Gabor filter output and therefore lead to a significantly improved recognition accuracy;

4. the combination method of local classifiers, which employs PCA along with FLD analysis, should be able to extract discriminating information from outputs of the local classifiers.

5. the implemented efficient facial expression system based on the proposed framework should be stable to process any given facial images by users and produce acceptable results.

This thesis focuses on studying facial expression recognition using fusion of a human vision system model and statistical approaches, especially on person-independent facial expression recognition from still images. Hence, spontaneous expression recognition based on video sequences is not considered in this thesis. Moreover, the applications of the methods in this study are limited to facial expression recognition from novel persons, which is considered to be extremely difficult and challenging in terms of substantial low recognition accuracy with conventional statistical approaches. It should also be noted that this thesis focuses on fusions of popular pattern recognition techniques, such as Gabor filters, local methods, PCA and FLD analysis; other techniques are beyond the scope of this study.

The organization of the thesis is as follows: Chapter 2, Chapter 3, and Chapter 4 deal with contour-based facial expression recognition algorithm, the modified HMAX model for facial expression recognition, and the composite orthonormal basis algorithm, respectively. In Chapter 5, a new framework is proposed for facial expression recognition, combining local Gabor features with classifier synthesis. The implementation and integration of the new framework is described in Chapter 6. And the thesis is concluded in Chapter 7 with a summary of the main contributions.

Chapter 2

Contour Based Facial Expression Recognition

As explained in Chapter 1, even though statistical approaches can extract features from gray-level images, there seems to be no evidence to demonstrate that these features are sufficient for facial expression recognition. Since the human ability to recognize expressions is known to be extraordinary, it is only natural to derive inspiration from empirical studies involving the human vision system. As mentioned in block-schematic (Figure 1.5) of Chapter 1, we will propose a contour based facial expression recognition algorithm, aiming at imitating the retinal ganglion cells of the human vision system. This is inspired by the fact that human retinal ganglion cells only signal the edges in a facial image, while face-selective cells in the inferior temporal area of the human brain respond maximally to these edge signals, according to [68]. A possible inference is that the contours of the face and of its components are *biologically plausible* features that play a key role in the human perception of facial expressions (see Figure 2.1 [67]). In many cases, it is observed that facial contours alone (as highlighted by the human beings' ability to appreciate cartoonists' sketches) do convey information that is adequate to recognize various expressions on the face, as evident from the human ability to understand and appreciate cartoons.

It is to be noted that a facial expression is not confined to a specific part of the face, and cannot be treated as a purely local phenomenon [66, 91]. As against this, some of the literature employs local features in a way that does not

Figure 2.1: Both natural images and cartoon images could clearly tell what the facial expression is [67].

include the spatial relationships existing, in general, among them; in other words, the local features are not treated *holistically*. However, it is common knowledge that the contours of a face act as a whole in conveying an expression. More specifically, the local contours around specific regions of the face (like the cheeks, mouth, eyes and eyebrows, and forehead) act together, i.e., globally, to compose an expression. This acts as a motivation for designing an expression recognition algorithm that deals with the local contours acting globally across the face. To this end, we propose an encoding mechanism that converts the facial contours to a grid-array (reminiscent of the human retinal cells around the fovea) that is input to a neural network with the property of self-organization (modeling a characteristic of the human brain) originally due to [42]. An interesting result is that the network generates a map, called the self-organizing map (SOM), that seems to exhibit distinct clusters for various expressions. This is a demonstration of the relevance of the extracted contours to facial expression recognition.

2.1 Contour Extraction and Self-Organizing Network

We consider the Japanese Female Facial Expression (JAFFE) [39] database, containing 213 images of 7 facial expressions of 10 Japanese female models, including 6 basic facial expressions (*happy, sad, angry, surprised, disgusted, scared*) and neutral faces [15]. The neutral expression can be treated as "no-expression". Since it has been found that the number of images in the JAFFE database is not

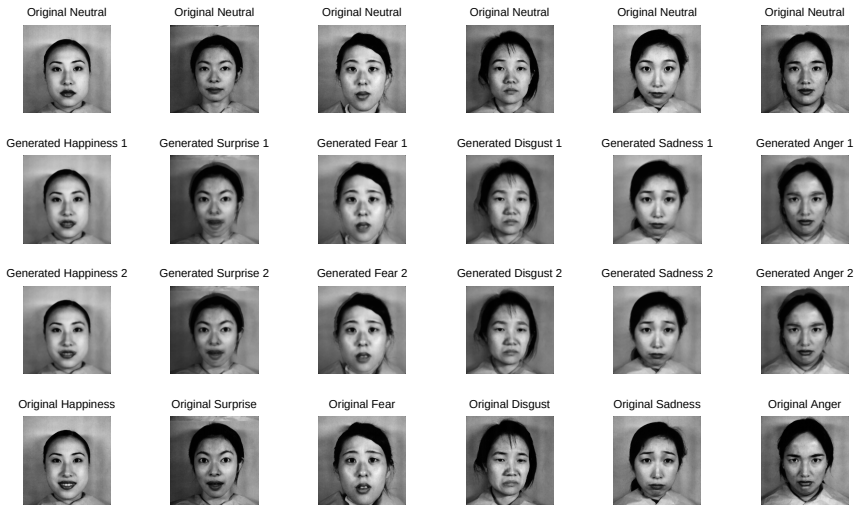


Figure 2.2: First row contains original images, while last row contains images of six basic expressions. Two rows in the middle consist of generated images.

representative enough for generating a self-organizing map (SOM) of the facial expressions (see Section 2.2.2 below for details), there arises the need to enlarge the database.

In order to generate new images of various expression, an example-based image synthesis algorithm [17] was adopted by using the neutral images as the references, and suitably interpolating between them and the existing expression images. As shown in Figure 2.2, the new images do appear to be visually different from the images of the JAFFE database. By a proper selection of these generated images, the JAFFE database was extended to 1080 images with 6 expressions of the same 10 female models. We crop the images to remove background information and normalize the size of images to 180×140. Finally, we apply a contour extraction algorithm to both extended database and original database.

For the extended database, we focus on the recognition of the 6 expressions. If we need to include the neutral expression in our recognition scheme, the training set should contain a sufficient number of images of neutral faces.

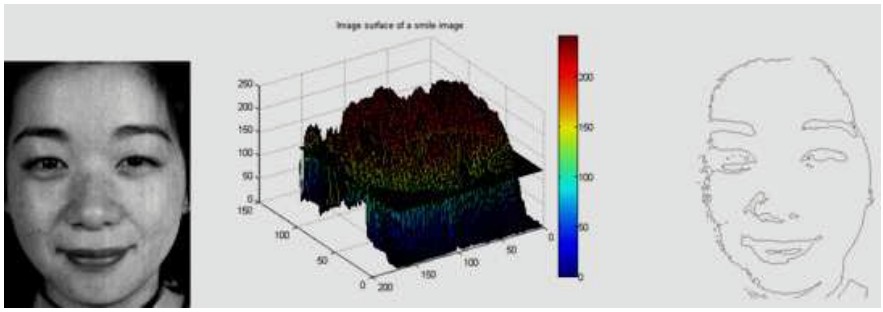


Figure 2.3: A smile image plotted as a surface where the height is its gray value. A plane intersects the surface at a given level and the resulting curve is a contour line of the original image.

2.1.1 Contour Extraction

The key to extracting contours appears to be the ability to distinguish between object contours and texture edges. Traditional edge detectors can be extended to suppress texture edges using local information around neighborhood of an edge, such as gradient of image intensity, anisotropic diffusion, and complementary analysis of boundaries and regions. We propose a new facial contour extraction algorithm based on level-sets.

Mathematically, a level set of a function $f : \Re^n \rightarrow \Re$ with n variables is a set of the form

$$(x_1, x_2, \cdots, x_n) \in \Re^n : f(x_1, x_2, \cdots, x_n) = c,$$

where c is a constant. In other words, it is the set where the function takes on a given constant value. When the number of variables is two, for example, when we deal with an image intensity function, $f(x, y)$, of spatial variables x (image height) and y (image width), with c specified as an integer between 0 and 255, the set $(x, y) : f(x, y) = c$ is called level curve or contour line. Figure 2.3 shows a smile image in the JAFFE database with the image surface of the smile image obtained using the method described above. By slicing the image surface using a plane $c = 120$, we generate a contour line which is a rough representation of facial contours. The main issue is whether we can obtain smooth and complete contours using such a method.

In the literature, a specific numerical algorithm [56] has been proposed to track contours and shapes in an image using partial differential equations (PDE) in order to arrive at smooth and complete contours. Traditional active contour algorithms, which use the level-set method [46], track only the zero level-set. First of all, an initial curve is embedded as the zero level-set of a given image surface. Secondly, the embedding curve is evolved according to a designed PDE. After the diffusion procedure, the initial curve approximately converges to certain desired object boundaries. However, human facial components are so complicated that the zero level-set turns out to be inadequate to represent any facial expression unambiguously. Therefore, we employ several level-sets together, invoking the histogram of the intermediate two-variable function which is a solution of the PDE.

A public MATLAB™ toolbox implementing level-set methods for image processing can be found in [70]. Unlike traditional active contour algorithms which use a given closed curve as the initial condition, a facial image itself is considered to be the level-set function in the proposed algorithm. The toolbox provides interfaces to solve the following partial differential equation (PDE):

$$\frac{\partial f}{\partial t} + \vec{V}_S \cdot \nabla f + V_N \left|\nabla f\right| = \beta \left|\nabla f\right| \tag{2.1}$$

where $\vec{V}_S$, V_N and β are forces (i) in the external vector field; (ii) in the normal direction to the curve; and (iii) based on the curvature of the curve, respectively.

Using the toolbox to extract contours, we assign image features to the forces in Equation 2.1 as follows:

- $\vec{V}_S = \nabla f = (G_x, G_y)$, where G_x,G_y are normalized gradient along x and y direction of the image respectively;
- $V_N = \frac{1}{(1+G_i)}$, where $G_i = |\nabla f| = \sqrt{G_x^2 + G_y^2}$;
- $\beta = \nabla \cdot \frac{\nabla f}{|\nabla f|} = \nabla \cdot \frac{\vec{V}_S}{G_i}$, the curvature of the edge map.

Notice that gradient vectors contain edge information of an image. Therefore, $\vec{V}_S$, the external vector, can pull curves towards the edges; V_N, the force in the normal direction of the edge, can stop curves around edges; and, the parameter,

β, helps to smooth the image and minimize the effect of noise. For details, see Osher and Sethian [56].

Since the PDE governs the dynamics of the image function, all the level-sets of f(x,y) evolve, and different level-sets represent contours of different facial components (See Figure 2.4). After the curve evolution, we diffuse the image such that flat areas are smoothed out, and edges are preserved and sharpened. Then we slice equally the PDE-solution surface whose height is gray value between 0 and 255 to obtain level-sets contours. For an unambiguous representation of various expressions, it seems to be difficult to indicate the number of level-sets to be used. From Figure 2.6 below, we observe that too small a number (less than 2) of level-sets lead to incomplete contours, while too large a number of (more than 6) level-sets lead to redundant contours. On the basis of the extensive experiments conducted, it has been found that 4 level-set contours represent facial expressions satisfactorily (see Section 2.2.4). For instance, Figure 2.5 shows the extracted contours which are found to provide sufficient information for facial expression recognition. The 4-level contours can be used either as a vector-set or in combination as a single set of contours. It is not yet known whether a vector-set representation of the level-set contours leads to better accuracy in the classification of expressions. For the purposes of the thesis, we plot all the contours of 4 levels together, and give the gradient-based edge strength for comparison.

2.1.2 Radial Encoding Strategy

After extracting facial contours from images, there is a need to efficiently encode them to form feature-arrays that are input to a neural network. Notice that one of the elegant properties of human vision system is the invariance to certain spatial transformation [61]. Moreover, it has been shown in [23] that radial encoded pattern is invariant to some transformations such as shift, scaling and (moderate) rotation. Therefore, we apply the radial encoding strategy to extracted contours, and the encoding procedure is as follows:

1. Place a radial grid on the feature (i.e., contour) image of the facial image under study (see in Figure 2.7);

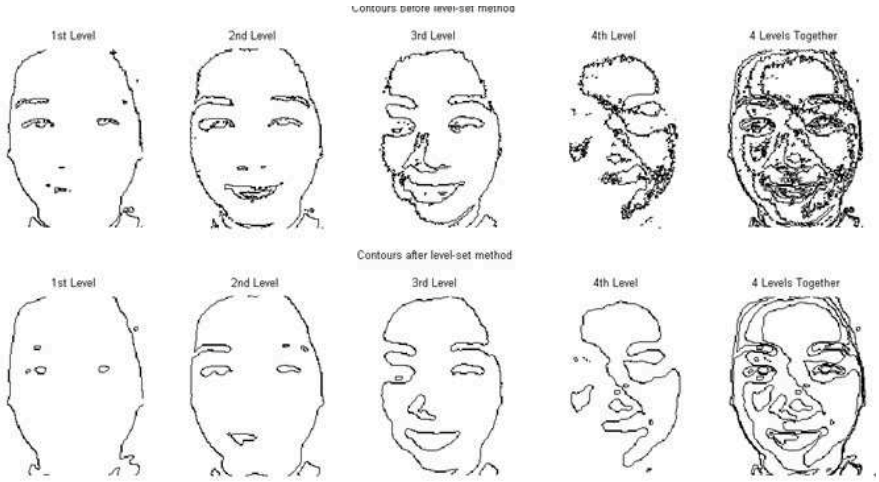


Figure 2.4: Contour results of the proposed algorithm. The first row contains contours obtained before smoothing and the second row contains contours obtained after smoothing. The first 4 columns contain results of 4 different levels while in the last column contours of all the 4 levels are plotted together.

2. Fix the center (x,y) of the radial grid at the center of the contour image (which is roughly the tip of nose as found on the contour);

3. Choose the radius r of the outermost circle of the radial grid according to $r = \frac{min(w,h)}{2}$, where w and h are the width and height of the contour image respectively;

4. Divide the radial grid into several sectors according to the grid resolution: angular vs. radius (i.e., resolution 30×12 will lead to 360 sectors);

5. Count the number of points that fall into each sector of grid, and assign it to the corresponding element in the grid-array.

2.1.3 Self-Organizing Network (SON)

Figure 2.8 shows the structure of a SON. Each neuron is represented by a d-dimensional weight-vector, $\vec{W}_i$ for the i^{th} neuron, where d is equal to the dimension of the input vector. Neurons are connected to adjacent neurons by a neighborhood relation that characterizes the topology of the network.

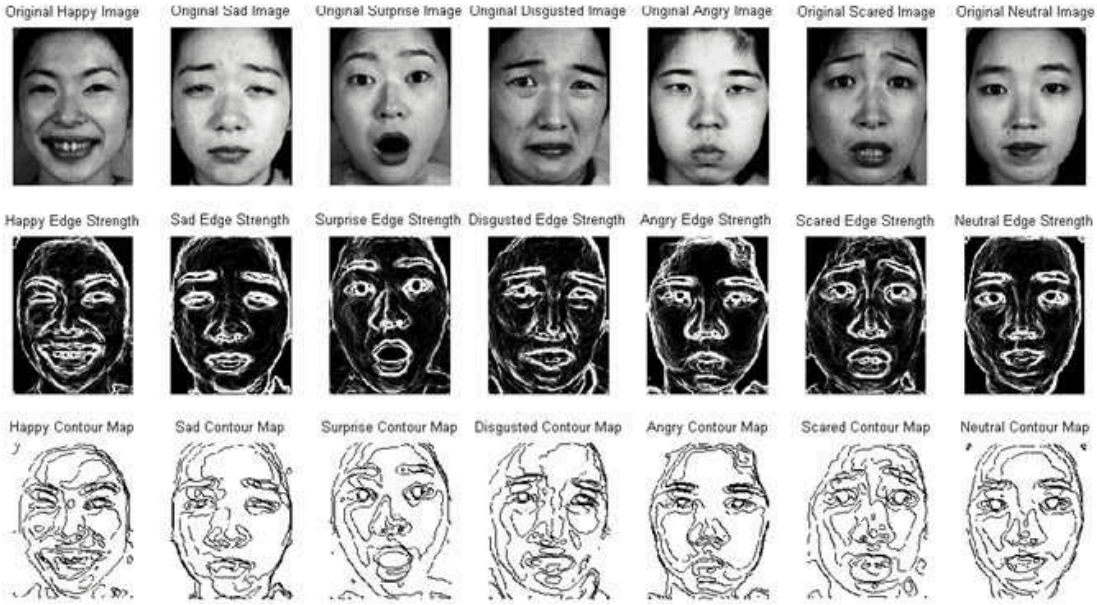


Figure 2.5: Gray-level images are in the first row, while edge strengths and level-set contours are in the second and third row respectively. Different columns contain images of different expressions. From the extracted contours, one can identify what the expression is.

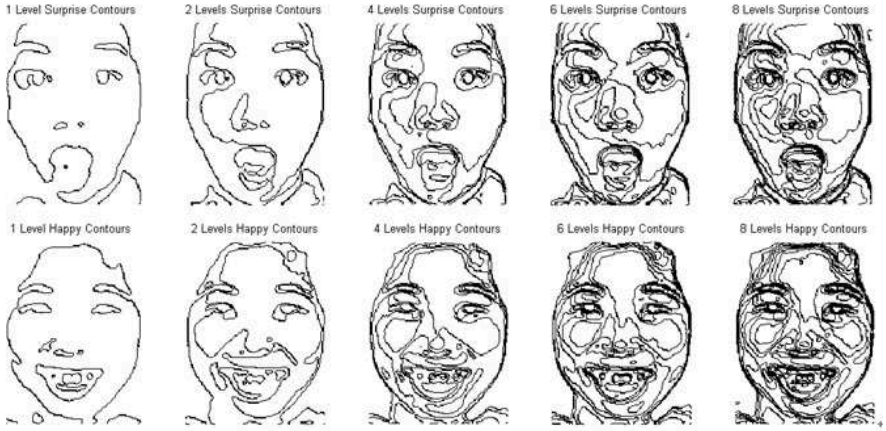


Figure 2.6: Different columns contain contour maps with different levels together.

By adopting a strategy originally proposed by [42], the SON maps complex relationships that may exist among high-dimensional input patterns into a two-dimensional pattern. The network is trained iteratively as follows:

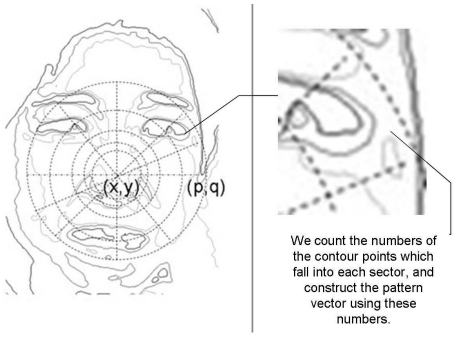


Figure 2.7: Radial grid encoding strategy. Central region has high resolution while peripheral region has low resolution.

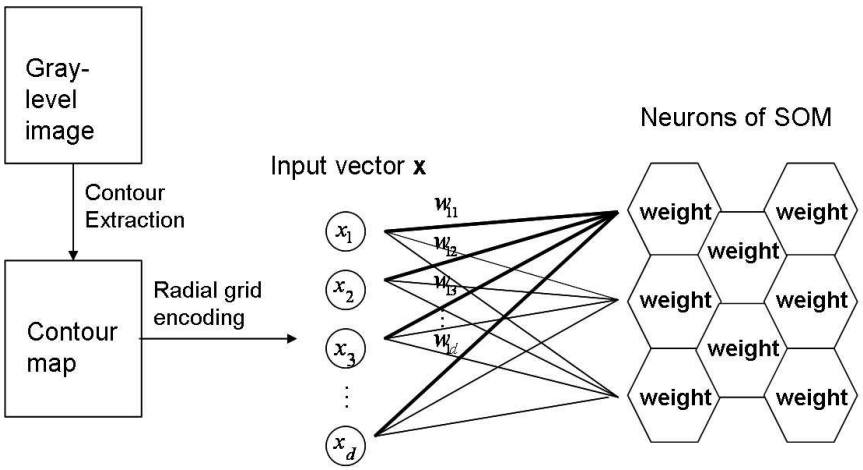


Figure 2.8: The structure of proposed network.

1. Randomly select one sample vector from the input data set, and calculate the Euclidean distances between it and all the weight-vectors $\vec{W}_i$ of the network.

2. Find the best matching unit (BMU), whose weight-vector is closest to the input vector. Call this neuron, c.

3. Update the weight-vectors of the network, as specified in Equation 2.2 below, such that the BMU is moved closer to the input vector.

4. Go to steps 2 and 3, repeat until there are no significant changes while updating weight-vectors.

$$\vec{W}_i(t+1) = \vec{W}_i(t) + \alpha(t)h_{ci}(t)[\vec{x}(t) - \vec{W}_i(t)] \tag{2.2}$$

where

$$\alpha(t) = \alpha_0 \exp\left(-\frac{t}{\tau_1}\right),$$

$$h_{ci}(t) = \exp\left(-\frac{d_{ci}^2}{2\sigma^2}\right),$$

$$\sigma = \sigma_0 \exp\left(-\frac{t}{\tau_2}\right),$$

t denotes the number of iteration; i, the i^{th} neuron; $\vec{x}(t)$, the input vector at iteration t; $\alpha(t)$, the learning rate at iteration t; $h_{ci}(t)$, the neighborhood kernel around the c; d_{ci}, the distance between unit c and i; and α_0,σ_0,τ_1 and τ_2 are initial parameters given by users. The neighborhood kernel is a non-increasing function of iteration at a distance of i from the c. Therefore, the topology of the network, in terms of neighboring connection of neurons, plays an important role in updating weight-vectors. It is found beneficial to have neighborhoods of a BMU at uniform distances. So in our implementation, we arrange neurons to form a hexagonal map.

After training, it is significant to note that the output map *does* become ordered, *i.e.*, the neurons are grouped into clusters according to the input expressions. The network has generated a map called the self-organizing map (SOM)

for expressions. An important byproduct of this result is that the chosen contours constitute meaningful features for recognition of expressions. In fact, the map can itself be used to classify facial expressions directly (see Section 2.2.2). However, more efficient neural architectures/algorithms like multi-layer perceptron (MLP) and support vector machine (SVM) can be employed for recognition of expressions from contours.

2.2 Simulation Results

We encode the contour images using a grid (radial vs angular) array of size 12×30. And we have also experimented with a lower grid resolution array 12×8 and a higher grid resolution array of 15×48. The experimental results (see Table 2.1) indicate that a lower resolution leads to a lower recognition accuracy, whereas higher resolution increases the computational load of the network without significantly improving recognition accuracy. This implies the need for a trade-off between computational complexity and accuracy.

In order to show that the encoded contour vectors contain enough information for recognizing facial expressions, we train a SON and check the homogeneity of encoded contour vectors in terms of the formation of clusters in SOM.

2.2.1 Checking Homogeneity of Encoded Expressions using SOM

The SON is implemented in MATLAB™ with Windows XP platform, and all simulations are conducted using a 2.60 GHz Pentium processor and 1 GB memory. The neural network is of size 70×70, and the dimension of the weight vector of each neuron is 360 (corresponding to the 12×30 grid-array). For training, we feed the arrays to the network as described in Section 2.1.3. Here, we investigate the homogeneity of the radial encoded contours, so all the images in the extended JAFFE database are used for generating the arrays for training SON. After training, the labels that indicate classes of facial expressions are given to the neurons as follows. For each individual neuron, the Euclidian distances between the neuron and all the input patterns are calculated, and the neuron is labeled by the

class of the closest input pattern. Then, the labels are plotted out to check the clusters of different classes of facial expressions. The time taken for generating an SOM from encoded contours is about 8 hours.

Figure 2.9 shows that the SOM obtained by the training operation *does* exhibit homogeneity of input patterns. Limitations of hardware do not permit experiments with larger SOMs. Still, the present SOM for the chosen network size (of 70×70) does admirably justify that the contours (i) do constitute features of relevance to recognition of expressions; and (ii) are detailed enough to recognize facial expressions. Based on these experiments, it can be concluded that the encoded contour vectors can be used to classify different facial expressions.

2.2.2 Encoded Expression Recognition Using SOM

In order to check the expression recognition accuracy of SOM, we need to separate database into training set and test set, and re-train the SON using the training set. After training, we compute recognition accuracy as follows:

1. Choose a novel sample from test set (N_s samples in total) as input to SON, and calculate Euclidean distances between input and all neurons.

2. Select the neuron with the shortest distance as the one that maximally responds to input. If the label of the winner neuron is the same as the label of input, increase the number of correctly matching samples, N_c.

3. Compute recognition accuracy: $Rate = N_c/N_s \times 100\%$.

In our implementation, the recognition accuracy is determined by a cross-validation experiment. There are generally two kinds of cross-validation: random cross-validation and person-independent cross-validation (also called "leaving-one-person-out"). The main difference is as follows: the first method randomly divides the database into several segments while the second one divides the database according to the number of persons in the database such that each segment contains all images belonging to only one person. After partition, one of these segments is picked out as the test set, and remaining segments are used for training. The above procedure is repeated so that all the segments are used

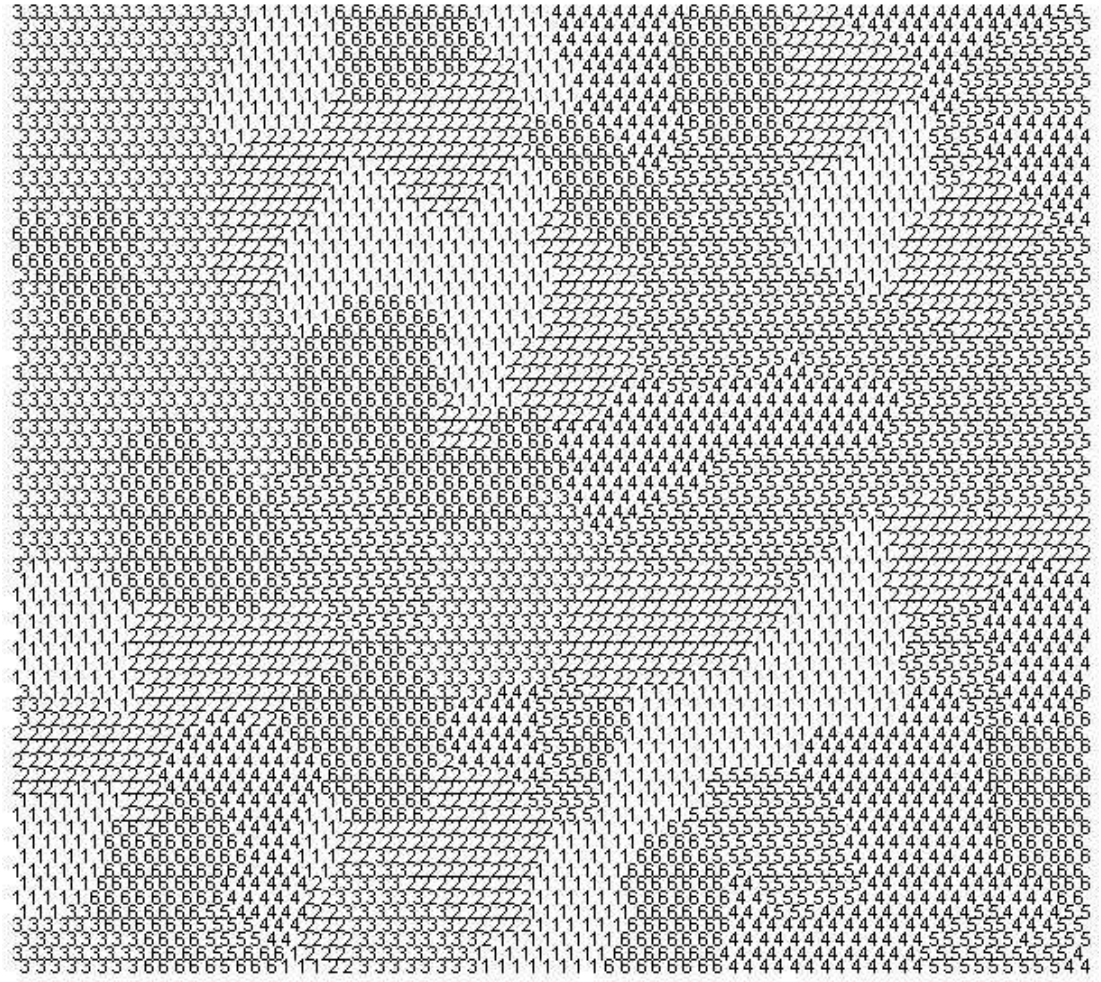


Figure 2.9: Labeled neurons of SOM with size of 70×70. Different labels, which indicate different expressions, are grouped in clusters. Labels from 1 to 6 indicate expressions of happy, sad, surprise, angry, disgusted and scared, respectively.

Table 2.1: Classification accuracies (%) of SOM with different sizes. The first row contains results of SOM using extended JAFFE database whereas the second row consists of results using original JAFFE database. Last two columns contain results of SOM with size of 70×70, of which input patterns are encoded under different resolutions. (L) stands for low resolution and (H) stands for high resolution. There are 972 images of 6 expressions for training in the extended (Ext.) JAFFE database and 120 images of 6 expressions for training in the original (Org.) JAFFE database.

DB.	30×30	50×50	70×70	70×70(L)	70×70(H)
Ext.	45.28	53.23	**58.11**	40.30	57.91
Org.	15.21	25.72	30.02	-	-

once as the test set, and the recognition accuracy is averaged over all the distinct segments.

Here we focus on the person-independent method because recognizing the expression of novel expressers is known to be difficult.

The time taken for testing SOM is 1.038 seconds. In addition, we also check the SOM results based on the original JAFFE database. The classification accuracies for both the extended and original databases are tabulated in Table 2.1, from which we observe the following: with original database, recognition accuracy is poor (maximum = 30.02%) because the training set is so small that generalization of SOM cannot be achieved. This leads to the need of using other classifiers that are valid for a small database (i.e., the original JAFFE database).

2.2.3 Expression Recognition using Other Classifiers

The other classifiers are multi-layer perceptron (MLP) with one hidden layer and k-nearest neighbor (KNN) with k setting to be one. The resolution of the radial grid is fixed to 12×30. Taking into account the limited database (213 images in total), we find that the dimension of the corresponding input vector to classifiers (360) is too high. Therefore, before we feed the encoded contour vectors to the classifiers, we first reduce their dimension using PCA. About 150 principal components are selected of which the cumulative energy is above 95%. Then these principal components are further processed using FLD analysis such

Table 2.2: Classification accuracy (%) of MLP and KNN based on the extended JAFFE. The first row gives results based on contour-based vectors, and the second row contains the results of image-based vectors. (R) indicates random cross-validation while (ID) means person-independent cross-validation (see Section 2.2.2).

Features	MLP(R)	KNN(R)	MLP(ID)	KNN(ID)
Contour	93.33	**96.67**	**66.67**	61.10
Image	77.82	84.26	54.08	53.42

that they are projected into $(C-1)$ discriminant features, where C is the number of classes. Finally after variance normalization, the projected Fisher features are fed into the MLP and KNN.

For comparison, we also apply PCA and FLD analysis directly to the facial images in the database, and obtain image-based feature vectors. Then we feed these vectors (also after variance normalization) into MLP and KNN. The training algorithm used in our implementation of MLP is the scaled conjugate gradient optimization [54]. The number of hidden neurons is based on the results of various experiments.

The MLP and KNN are used to classify the contour-based vectors and image-based vectors obtained from the extended JAFFE database (having a total number of 1080 images of 6 expressions). The results are shown in Table 2.2, from which we observe that contour-based features produce higher recognition accuracies irrespective of the cross-validation method employed. Then, in order to compare with results reported in the literature, we also apply MLP and KNN to the contour-based vectors and image-based vectors obtained from the original JAFFE (having a total number of 213 images of **7** expressions, including the neutral face). The experimental results in Table 2.3 show that the contour-based algorithm also works well on the original JAFFE database. Moreover, this result confirms that the poor recognition accuracy of SON based on the original JAFFE database is due to the limited size of the training set.

In order to further test the capability of the proposed contour-based algorithm, it is applied to the Taiwanese Facial Expression Image Database (TFEID)

Table 2.3: Classification accuracy (%) of MLP and KNN based on the original JAFFE database. The first row gives results based on contour-based vectors, and the second row contains the results of image-based vectors. (R) indicates random cross-validation while (ID) means person-independent cross-validation (see Section 2.2.2).

Features	MLP(R)	KNN(R)	MLP(ID)	KNN(ID)
Contour	**90.92**	82.386	**66.05**	59.77
Image	74.76	79.52	55.72	49.58

[8], which contains facial expression images of 40 persons (20 males and 20 females). Each person has 8 images corresponding to 8 expressions: happy, sad, surprise, disgusted, scared, neutral and contempt. In the simulation, we exclude the contempt expression, and focus on the 7 basic expressions, including the neutral face. It has been found that the TFEID database is incomplete: for some particular persons in the database, some of the expression images are missing. Therefore, we use 268 images in TFEID database for the simulations. All the images are cropped such that background information is removed and then the size of images is normalized to be uniform (180×140). The resolution of radial grid is kept to be 30×12 and parameters of KNN and MLP are also kept the same as those of JAFFE.

Table 2.4 shows the recognition accuracies of the proposed algorithm on different level-sets contours from the JAFFE and TFEID databases using person-independent cross-validation. It is clear that the proposed algorithm works well for both JAFFE and TFEID databases. In addition, Table 2.4 contains results with respect to contours with different level-sets, confirming that the contours of 4 level-sets together lead to satisfactory expression recognition.

2.2.4 Human Behavior Experiment

We have invited 15 naive subjects (called "participators"), having normal expression judgement, to recognize the facial expression images in the original JAFFE database. Using the newly developed software (Figure 2.10), we show 213 images

Table 2.4: Classification accuracy (%) of MLP and KNN based on the original TFEID and JAFFE databases using person-independent cross-validation with respect to contours with different level-sets .

TFEID	1 Level	2 Levels	4 Levels	6 Levels	8 Levels
MLP	69.40	63.43	**85.45**	73.51	76.87
KNN	67.91	67.16	79.10	79.10	79.48
JAFFE	1 Level	2 Levels	4 Levels	6 Levels	8 Levels
MLP	48.36	54.46	**66.05**	57.28	50.23
KNN	46.95	52.11	59.77	51.17	51.17

Table 2.5: Classification accuracies (%) of different expressers. The first row gives results based on human behavior, and the second row contains the results of MLP using the proposed algorithm. Column 2 to column 11 is for ten expressers (here the order of expressers is the same as the one in the original JAFFE) respectively while the last column is the average value.

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	Average
Human	77.02	61.04	68.54	55.36	78.58	62.25	56.43	66.67	49.66	81.18	65.67
MLP	56.52	68.18	77.27	50.00	66.67	76.19	60.00	52.38	71.43	81.82	66.05

of **7** expressions sequentially to the participators who are allowed enough time to make a judgement on the expression of the image, and whose classification of expressions is automatically recorded by the software. The average recognition accuracy over 15 subjects is 65.67%. An interesting phenomenon is that participators report that the expressions of some persons are difficult to identify, leading to differences of recognition accuracies among different persons. The results of both human and and MLP-based automatic recognition (using the proposed algorithm and person-independent cross-validation) of expressions of ten persons are tabulated in Table 2.5 which shows that performance of the proposed algorithm is close to human perception.

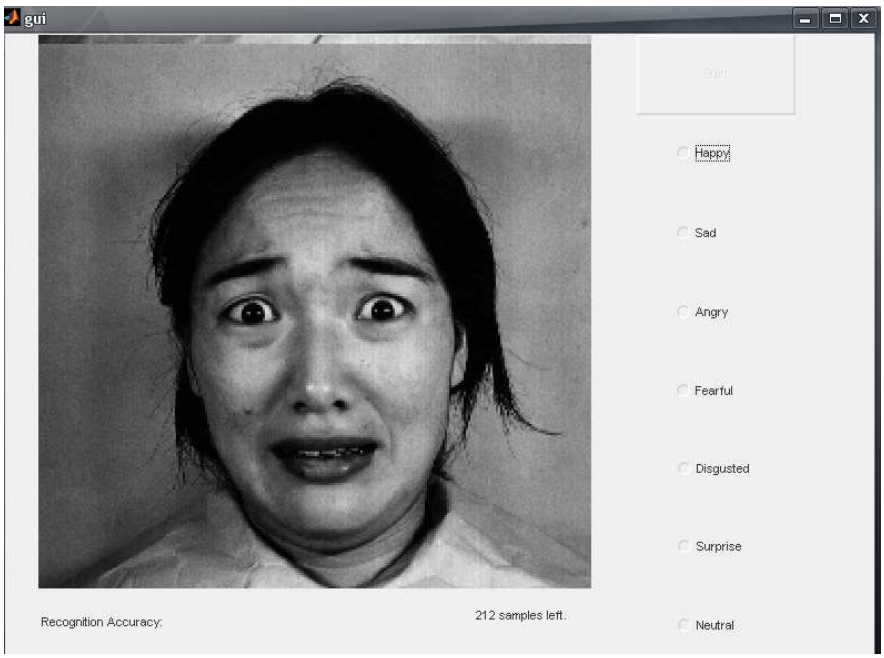


Figure 2.10: Snapshot of the user interface for human to recognize expressions using the JAFFE database.

2.3 Discussions

In this section, we first compare the contour-based algorithm (and the best results obtained by MLP on the JAFFE database) with others, as applied to the JAFFE database in the literature.

Zhang et.al. [91] report a result of 90.1% for 6 expressions (neutral expression is excluded) using random cross-validation. Huang et.al. [32] report a result of 51% for 7 expressions using person-independent cross-validation. Both Lyons et.al. [52] and Shinohara and Otsu [66] use person-independent cross-validation to get 75% and 69.4% recognition accuracies, respectively. However, there are only 193 images of 9 subjects in their experiments. It is not clear which expresser has been excluded. Interestingly, Table 2.5 implies that exclusion of certain expresser could be crucial when training the network. In [85], for 183 images of 10 subjects, excluding the neutral expression, both cross-validation methods have been used, and the reported results are 98.36% and 77.05%, respectively. Moreover, as in [52], their algorithm required additional information (such as manually chosen

geometric points and semantic expression ratings) for recognizing expressions. In the light of the above facts, it can be concluded that the contour-based algorithm is comparable to those in the literature.

2.4 Summary

In this chapter, motivated by the human vision system, and invoking some aspects of the known neural structure (with respect to its retina and the visual cortex), we have demonstrated, perhaps for the first time, that the contours of the facial components, considered as a *whole*, can be successfully employed for automatic facial expression recognition. Experimental results have indicate that radial encoded contours are biologically plausible features for facial expression recognition since the contour-based algorithm has close performance to that of human beings.

Chapter 3

Modified HMAX for Facial Expression Recognition

In Chapter 2, a contour-based algorithm has been proposed based on the retinal ganglion cells, which are the initial stage of human vision system. Although the performance of contour-based algorithm on recognizing expressions is close to that of human beings, it still needs to be improved in terms of recognition accuracy. To this end, we now focus on considering more advanced processing stages in human vision system: V1 cells and face-selective cells. As explained in Chapter 1, standard HMAX, modeling V1 cells using Gabor filters and incorporating hierarchical structure of human vision system, has been proposed for rapid object recognition. However, HMAX fails to recognize expressions because it lacks the ability of face processing. In this chapter, we will modify HMAX model according to the recent biological findings about the face-selective cells in FFA, such that the modified HMAX can handle facial expression recognition (see Figure 1.5, the block-schematic in Chapter 1).

3.1 HMAX with Facial Expression Processing Units

A simple way to extend HMAX to expression recognition is to add the face processing layer. Since we do not know the exact biological mechanism to process

face images, we employ, in its place, some statistical approaches: the output of C2 units is subjected to PCA and FLD analysis in order to obtain discriminating features for facial expression recognition. For the extended architecture of the HMAX to process face images to recognize expressions, see Figure 3.1. The detailed procedures are as follows:

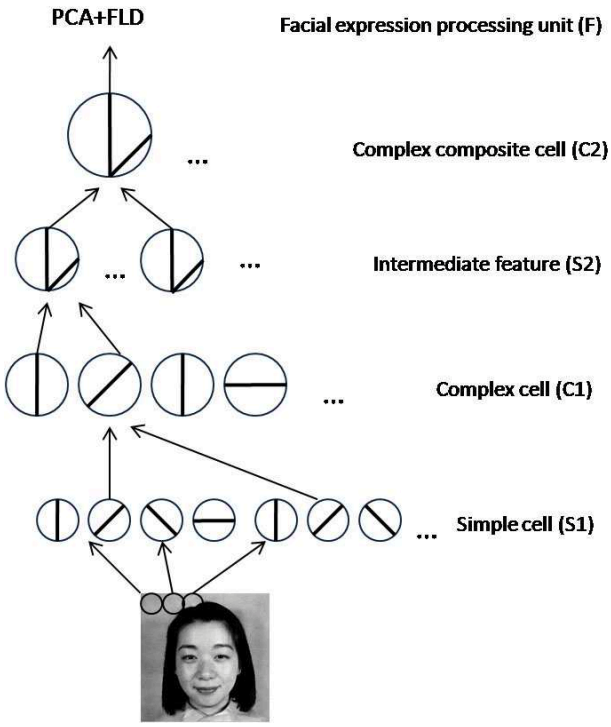


Figure 3.1: Structure of HMAX with facial expression processing units.

1. The S1 responses are first obtained by applying a set of Gabor filters to the input image I. Recall that the Gabor filter can be described by the following equation:

$$F(x,y) = \exp(-\frac{(\hat{x}^2 + \gamma^2\hat{y}^2)}{2\sigma^2}) \times \cos(\frac{2\pi}{\lambda}\hat{x}), \tag{3.1}$$

where

$$\hat{x} = x\cos\theta + y\sin\theta, \hat{y} = -x\sin\theta + y\cos\theta \tag{3.2}$$

Here, (x,y) refers to the 2D coordinate system of the input image. The five parameters (orientation θ, aspect ratio γ, effective width σ, phase ϕ

and wavelength λ) determine the properties of the Gabor output. Since the parameters ϕ and γ have little influence on the performance of the recognition system, ϕ is set to 0 and γ, to 0.5. Four orientations ($\theta = 0°, 45°, 90°$ and $135°$) have been found to be sufficient for our purposes. The remaining parameters σ and λ are determined by the following equations based on the tuning properties of cortical cells according to [64]:

$$\sigma = 0.0036 \times \mathcal{S}^2 + 0.35 \times \mathcal{S} + 0.18 \tag{3.3}$$

$$\lambda = \frac{\sigma}{0.8} \tag{3.4}$$

where $\mathcal{S}$, the filter size to determine σ and λ, varies from 7 to 39 in steps of two, according to [64].

2. C1 units pool responses over S1 units using max operation, and have some tolerances to certain moderate shift and scale changes.

3. S2 layer contains RBF-like units that are tuned to object-parts and compute a function of the distance between the input units and the stored prototypes.

4. C2 units perform a max operation over the whole visual field and provide the intermediate encoding of the stimulus. The difference between standard HMAX and HMAX with feature learning lies in the connectivity from C1 to S2 layer: in standard HMAX, these connections are hard-wired to generate 256 combinations (if size of 2×2) of C1 inputs while, in HMAX with feature learning, the prototypes are learned from the training set.

5. Facial expression processing units, F, take the responses of C2 units as input, and perform PCA and FLD to extract discriminating features which act as inputs to the classifier to recognize facial expressions.

Simulation results demonstrate that the F units contribute significantly to improving facial expression recognition, when compared to the original HMAX (see Section 3.4.1 for details). Therefore, these F units are always used in other modified versions of HMAX also, in the following sections.

3.2 HMAX with Hebbian Learning

The improvement effected by the facial expression processing units in recognizing expressions is not satisfactory, especially for cross-database task [1] (see Section 3.4.1). Even using the HMAX with RBF-like feature learning, experimental results (see Section 3.4.2) show that recognition accuracies of HMAX trained by natural images on cross-database task are not good enough. The conclusion is that, when natural images are used for training, the RBF-like feature learning strategy cannot capture universal facial expression information. Furthermore, when the HMAX is trained on facial images, expression recognition accuracies on test data from individual databases are satisfactory but those on test data from cross databases show instability.

Here it is to be noted that the JAFFE database contains more variations (such as pose and illumination changes) than the TFEID database. Therefore, when the TFEID database is used as a training set, the prototypes extracted may be not suitable for facial images in the JAFFE database, thereby indicating that, in order to achieve good generalization, the learning stage needs training samples with large variations.

To overcome the limitations of the above learning strategy, we now propose a Hebbian learning [14] strategy to generate prototypes from C1 to S2 layer. Let $C1_i$ denote the outputs of C1 layer, where i $(= 1, 2, 3, 4)$ stands for orientations: for every element of $C1_i$ in the same position (x, y), we compute the S2 response using the following formula:

$$S2(x, y) = \phi(\sum_{i=1}^{4} w_i C1_i(x, y)) \tag{3.5}$$

where $\phi(\cdot)$ is the Gaussian-like tuning function and w_i are the weights for 4 orientations. The weights are learned in a Hebbian learning manner:

$$\bar{\mathbf{W}}^{new} = \bar{\mathbf{W}}^{old} + \alpha \bar{\mathbf{S2}}(\bar{\mathbf{C1}} - \bar{\mathbf{W}}^{old}) \tag{3.6}$$

where $\bar{\mathbf{S2}}$ is the output of S2 units; $\bar{\mathbf{C1}}$ is the output of C1 units; and α is the learning rate and is fixed to 0.01 in our implementation. After training with facial

[1] Refer to Section 3.4 for descriptions of (i) test strategy and (ii) database.

images, the linear combination of 4 orientations can be used to represent a facial expression information. In the next stage, all the S2 outputs are fed into facial expression processing units without performing max operation. This procedure is to keep as much information as possible for the subsequent processing so that the system can deal with databases with different degrees of variations. Experimental results (see Section 3.4.2) show the improvement of Hebbian learning on cross-database task.

3.3 HMAX with Local Method

The HMAX and its modifications, as described above, are holistical, i.e., they take a full image as input. A serious limitation of such methods is that local information relevant to facial expressions may be lost. Recall that face processing cells in FFA capture local facial information in each cell acting as a local receptive field and possibly reconstruct a face, preserving most facial information, by combining local information [76, 77, 78]. Furthermore, local methods have shown promising results not only in facial expression recognition but also in other visual recognition tasks [5]. We now consider combining the HMAX model with local methods. See Figure 3.2 for such an architecture. The corresponding procedure is as follows:

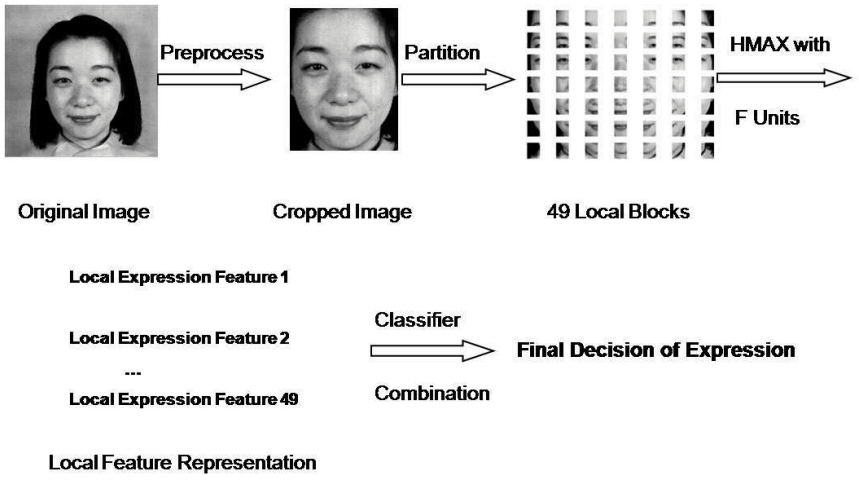


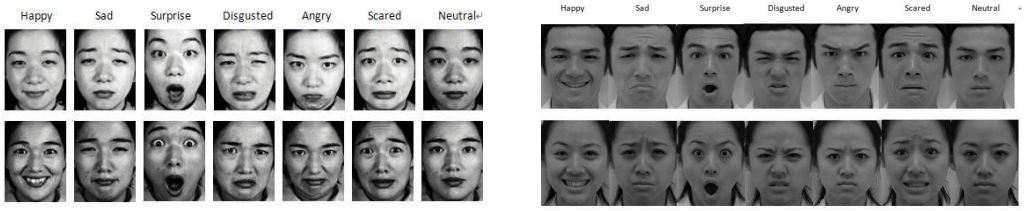
Figure 3.2: Sketch of the HMAX model with local methods.

1. After cropping images such that the background information is removed and the size of input images is uniform, each image is divided into several local regions with overlap half of their size. The facial components in the local regions should be as complete as possible while the local regions should be small enough such that local features can be extracted from the facial components. To this end, 49 local blocks are used to achieve a proper trade-off between the locality and the completeness of the facial components.

2. The original HMAX with facial expression processing units is applied to every local blocks of one facial image to obtain local features. Since the local block, containing local facial information, is small, the hard-wired 2 by 2 arrangements of four orientations are sufficient for extracting local features.

3. A set of local classifiers, such as k-nearest neighbor (KNN), is applied to the extracted local features to make local decisions. The outputs of local classifiers can be used as the intermediate features and their combination can lead to a global decision for recognizing expressions.

Classical combination rules, such as Borda count [48] and decision template [49], can be used to combine local classifiers to obtain global decision. However, the Borda count, based on voting, does not utilize information in training data. On the other hand, the decision template, which is a nearest-mean classifier in the decision space, may not capture discriminating information for high dimensional decision space. Therefore, it has been found expedient to concatenate outputs of all local classifiers for one facial expression image to generate the intermediate feature matrix of that image. In this manner, every facial expression image is represented by an intermediate feature vector. PCA and FLD analysis are then used to project the intermediate features into a discriminating low-dimensional subspace which can be effectively classified. Let C be the number of classes and Nc be the number of extracted features. Note the FLD analysis can at most extract $Nc = C - 1$ discriminating components from the input data, which may be insufficient to represent the global features with high complexity. Therefore, we adopt the recursive FLD (or RFLD) analysis [86], which uses the basic idea of



(a) Cropped gray-level images in the JAFFE database.

(b) Cropped gray-level images in the TFEID database

Figure 3.3: Samples in the two facial expression databases.

FLD analysis, but extracts one feature component at each iteration, and discards the information already extracted by previous iterations from all the samples before the next iteration. Further, in order to avoid over-fitting, we invoke the regularization method in RFLD analysis as follows:

$$S_W \rightarrow S_W + \beta \cdot Ave(Eigv(S_W)) \cdot I, \tag{3.7}$$

where β is the regularization factor which controls the influence of the regularization term; S_W is the within-classes scatter-matrix; $Ave\ (Eigv(S_W))$ denotes the average eigenvalue of S_W, and I is the identity matrix. In this manner, we control the final performance of HMAX (combined with local methods) by the two parameters, β and N_c . We study their effect on the final recognition performance in Section 3.4.3.

3.4 Simulation Results

As in Chapter 2, we use the following facial expression databases for experiments: (1) JAFFE database; (2) TFEID database. Figure 3.3 shows some samples in the two databases. All the images are cropped to remove background information, and normalized to uniform size (180×140).

There are, in general, two kinds of methods for testing the performance of the modified HMAX:

- Individual database test: The training images and testing images are from the same database. Since both the JAFFE and TFEID databases con-

Table 3.1: Recognition results (%) on individual database task.

	Standard HMAX	HMAX with F units
JAFFE	32.39	39.44
TFEID	44.03	64.93

tain limited samples, we employ the leaving-one-person-out cross-validation strategies.

- Cross-database test: The training images are from one database while the test images are from another database in turn: 1) use the JAFFE database for training and the TFEID database for testing; 2) vice versa. The goal is to check whether the expression features, extracted by a recognition system, are representative enough such that a new facial expression image from another database can also be recognized.

3.4.1 Experiments Using HMAX with Facial Expression Processing Units

HMAX with facial expression processing units is applied to both the JAFFE and TFEID databases to obtain feature vectors. Then the nearest-neighbor classifier is used to classify different expressions. As described above, both individual database test and cross database test are performed in the simulation. Table. 3.1 shows the recognition results on individual database recognition; and Table. 3.2, the recognition results on cross database recognition. We observe that HMAX with facial expression processing units outperforms the standard HMAX. However the performance is still not satisfactory. This confirms that the hard-wired feature prototypes in the standard HMAX are not suitable for facial expression recognition.

Table 3.2: Recognition results (%) on cross database task.

	Standard HMAX	HMAX with F units
JAFFE train, TFEID test	19.78	27.61
TFEID train, JAFFE test	17.37	24.41

Table 3.3: Recognition results (%) of HMAX with Hebbian learning.

JAFFE	TFEID	JAFFE train, TFEID test	TFEID train, JAFFE test
78.87	97.01	52.99	39.83

Table 3.4: Recognition results (%) of HMAX with RBF-like learning.

JAFFE	TFEID	JAFFE train, TFEID test	TFEID train,JAFFE test
77.46	96.27	60.19	29.80

3.4.2 Experiments Using HMAX with Hebbian Learning

HMAX with Hebbian learning is applied to both individual database test and cross database tests. The results of using nearest-neighbor classifier are presented in Table. 3.3. For comparison, HMAX with feature learning strategy is also applied to the same problem. We employ images from the JAFFE and TFEID databases to train the HMAX with feature learning. For the recognition results, see Table. 3.4. We observe that the results on individual database task are slightly higher than those of HMAX with RBF-like learning while the results on cross-database task are more stable than those of HMAX with RBF-like learning.

3.4.3 Experiments Using HMAX with Local Methods

HMAX with local methods is applied to both individual database test and cross database test. In the computational experiments, we vary β and Nc to obtain the best results. Table. 3.5 shows the recognition accuracies of HMAX with local methods on an individual database task; and Table. 3.6, the recognition accuracies of HMAX with local methods on a cross database task. The results of using decision template and Borda count to combine local classifiers are also given for comparison. It is obvious that HMAX with local methods can lead to

Table 3.5: Recognition results (%) of HMAX with local methods on individual database task.

	Decision Template	Borda Count	PCA + RFLD
JAFFE	75.12	73.24	79.81 ($\lambda = 1.1, Nc = 16$)
TFEID	96.27	96.27	98.88 ($\lambda = 1, Nc = 6$)

Table 3.6: Recognition results (%) of HMAX with local methods on cross database task.

	Decision Template	Borda Count	PCA + RFLD
JAFFE train, TFEID test	52.62	50.37	60.82 ($\lambda = 1.1, Nc = 30$)
TFEID train, JAFFE test	41.31	38.50	50.70 ($\lambda = 1.3, Nc = 25$)

satisfactory results for both individual database recognition and cross database recognition.

3.5 Summary

In this chapter, inspired by the recent biological findings about face-selective cells in FFA, we have modified HMAX model by adding face processing units and invoking local methods such that it is able to recognize facial expressions. Experimental results have demonstrated that local methods, combined with traditional statistical approaches, such as PCA and FLD analysis, significantly improve the performance of facial expression recognition.

Chapter 4

Composite Orthonormal Basis for Person-Independent Facial Expression Recognition

In Chapter 3, modified HMAX models, based on face-selective cells in FFA, have been proposed for recognizing expressions with improved performance on both individual database and cross-database tests. We have observed that most existing algorithms, including the proposed contour-based algorithm and modified HMAX model, seem to perform well when the identity of the person is known, i.e., when *person-dependent* expression recognition is considered. However, they are significantly less efficient for *person-independent* expression recognition. One of the reasons for this limitation is that there is a substantial overlap between the subspace of expressions and the subspace of identities.

In this chapter, we focus on improving performance of person-independent expression recognition. More explicitly, the unresolved and, hence, challenging, problem is automatic recognition of expression of a *stranger* (i.e., a face not in the database). Alternatively, we deal with the following cross database expression recognition problem: *Are the features of facial expression images from a given database, extracted by a recognition system, representative enough such that a new facial expression image from another database can also be recognized?*

To solve this problem, we will build an expression processor, imitating the expression-selective cells in STS of human vision system (see block-schematic in

Figure 1.5 of Chapter 1). Since the detailed mechanisms of expression-selective cells are yet unknown, we propose a statistical approach to present face images, by developing a **Composite Orthonormal Basis** (COB) from the given training set of face images of the (facial expression) database, and extracting from it an expression subspace with the identity information removed as much as possible. This subspace, corresponding to the global features of facial expression, enables the algorithm to extract universal expression features which are independent of identity information. Assuming that expression-selective cells have the similar properties (e.g, capturing local information) as face-selective cells in FFA, we enhance the power of the COB (to classify expressions) by combining it with local methods. More specifically, we employ the COB to represent local blocks of the input face images, thereby obtaining local expression features, which are then fed to a set of local classifiers. At the final stage, the local classifiers are combined to arrive at the final classification of expressions.

4.1 Composite Orthonormal Basis Algorithm

Let N^0 be the universal neutral which could be chosen as any one of the neutral faces from the training database; and N^k for $k = 1, 2, \cdots, n$, the other neutrals, where n is the number of persons in the database. Let their individual expressions be denoted by $\xi_{k,\ell}$ for $\ell = 1, 2, \cdots, C$, where k corresponds to the person's identity and $C = 7$ denotes the 7 basic expressions. Assuming that a neutral face of a person contains most of the identity information of that person, we separate expression from identity by subtracting the neutral face image from the corresponding expression image, and call the resultant the *flow-matrix* of the person for that expression or, more simply, *expression flow-matrix*, $F_{k,\ell} = N_k - \xi_{k,\ell}$, for $k = 1, 2, \cdots, n$ and $\ell = 1, 2, \cdots, C$.

Since the universal neutral and the neutrals of persons are not the same (except when a person's neutral has been designated as the universal neutral), the difference between the universal neutral and the neutral of each person is non-trivial. This difference is called the *neutral flow-matrix*, $F_k^0 = N^0 - N_k$, for $k = 1, 2, \cdots, n$, which is treated as a possible feature for the identity of person k. However, from experimental results shown in Section 4.2.1, we find that this

identity feature is still correlated with the above expression feature. One possible reason is that we are using only a linear algebraic operation on the face images. In an attempt to minimize this correlation for the chosen features, we propose a novel composite orthonormal basis for representing a face image in which the subspaces of expressions and identities are disjoint.

4.1.1 Composite Orthonormal Basis

Let $s1 \times s2$ be the size of an image, and let $s = s1 \times s2$. Note that each image can be represented as a stacked vector of dimension s. We assume that the flow-matrices have been transformed into vectors. The steps to generate the COB are as follows:

1. Compute the principal components (PCs) of $\{F_k^0 \mid k = 1, 2, \cdots, n\}$. Let these be denoted by P_r^0 for $r = 1, 2, \cdots, \rho_0$, where $\rho_0 \leq n$ is the number of chosen principal components for the neutral flow-matrices.

2. For each $l = 1, 2, \cdots, C$, compute the PCs of $\{F_{k,l} \mid k = 1, 2, \cdots, n\}$. Let these be denoted by $P_{l,r}$, for $r = 1, 2, \cdots, \rho_l$, where $\rho_l \leq n$ is the number of principal components chosen for the expression flow-matrices.

3. Apply the Gram-Schmidt procedure to generate the orthonormal composites:

   (a) Consider the first two sets of PCs: $\{P_1^0, P_2^0, \cdots, P_{\rho_0}^0\}$ and $\{P_{1,1}, P_{1,2}, \cdots, P_{1,\rho_1}\}$.

   (b) Bias each of the components of $P_{1,r}$ by $F_1 = F_{1avg}/\rho_1$ to get the new set of modified pseudo-PCs for expression 1: $\{P_{1,1} + F_1, P_{1,2} + F_1, \cdots, P_{1,\rho_1} + F_1\}$. Here F_{1avg} is the mean of $F_{k,1}$. And this operation is to compensate the mean which is subtracted in the principal components calculation.

   (c) Invoke the Gram-Schmidt procedure to generate, from $P_{1,1} + F_1$, a vector orthogonal to the set $\{P_r^0 \mid r = 1, 2, \cdots, \rho_0\}$, and normalize it to get $Q_{1,1}$.

   (d) Repeat the above step to generate, from $P_{1,2} + F_1$, a vector orthogonal to the set $\{P_r^0 \mid r = 1, 2, \cdots, \rho_0\}$ and $Q_{1,1}$. Normalize it to get $Q_{1,2}$.

(e) Repeat the above step for all the remaining modified pseudo-PCs of expression 1, and similarly for the other modified PCs of expression $2 - C$.

This eventually leads to the COB whose elements are given by $\{P_r^0 \mid r = 1, 2, \cdots, \rho_0\}$, $\{Q_{1,r} \mid r = 1, 2, \cdots, \rho_1\}$, $\cdots$, $\{Q_{C,r} \mid r = 1, 2, \cdots, \rho_C\}$. Recalling that N^0 is the universal neutral, and denoting the mean of F_k^0 by F_{avg}^0, we can express $\xi_{k,l}$, which is expression ℓ of person k, as a linear combination of the elements of the COB, N^0 and F_{avg}^0 as follows:

$$\xi_{k,l} = N^0 + F_{avg}^0 + \sum_{r=1}^{\rho_0} \alpha_{0,r} P_r^0 + \sum_{r=1}^{\rho_1} \alpha_{1,r} Q_{1,r} + \cdots + \sum_{r=C}^{\rho_C} \alpha_{C,r} Q_{C,r}, \tag{4.1}$$

where the unknown coefficients $\alpha_{s,r}$ are to be determined by solving linear algebraic equations. If the features chosen are appropriate, $\alpha_{\ell,r}$ refers to expression ℓ, and $\alpha_{0,r}$, to neutral faces. Interestingly, from the experimental results of Section 4.2.1, we find that these coefficients indeed exhibit homogeneity with respect to expressions. The intrinsic geometry of the facial surface as extracted from the flow-matrices $\{F_k^0 \mid k = 1, 2, \cdots, n\}$ can be attributed to neutral face, and the expression-based additional principal components correspond to the extrinsic geometry attributed to the C facial expressions.

4.1.2 Combination of COB and Local Methods

As mentioned earlier (in Chapter 1), face-selective cells in FFA capture local facial information in each cell acting as a local receptive field, and possibly reconstruct a face, preserving most facial information, by combining local information [76, 77, 78]. Assuming that expression-selective cells in STS have the same properties as face-selective cells in FFA, we combine the COB with local methods in order to improve its ability to recognize facial expressions.

The steps involved is described as follows. We mark manually the positions of the eyes and mouth, and hence the specific regions of interest in the image, while simultaneously achieving a uniform image size of 150×120. Then, we divide each image into several local regions, some of which contain (a) the corners of the eyebrow and mouth; and (b) the wrinkles. In our implementation, the

neighboring local regions are designed to overlap half of their size. The actual number of regions is chosen on the basis of the following considerations: 1) the facial components should be as complete as possible in the local regions; and 2) the local regions should be small enough such that local features could be extracted from the facial components. The satisfactory number of local blocks needed is determined using experimental results (see Section 4.2.2 below).

We represent every local region using the COB, and determine the coefficients which are then subjected to the FLD analysis in order to extract discriminating features based on the local expression coefficients. The resulting feature vectors are normalized to have zero mean and unit standard deviation, before using them as the input to a local classifier for each local feature. The local classifier we choose is the k-nearest neighbor (KNN) with $k = 1$. Its output is usually a scalar denoting the class label of the input data. Since we are now dealing with local features, we cannot expect a local classifier, based on local information from one local patch, to make a deterministic global decision on the correct expression. Therefore, we treat the elements of the C-dimensional vector of the KNN classifier output as probability estimates of the C classes. To this end, we modify the KNN algorithm as follows:

1. For a given test input vector, compute the Euclidean distances from the test input to all training vectors;

2. For each class, find the minimum distances d_i, where $i = 1, 2, \cdots, C$;

3. Compute the probabilities for each class using the following indication function:
$$P_i = \frac{(1 + d_i)^{-1}}{\sum_{k=1}^{C}(1 + d_k)^{-1}}, i = 1, 2, \cdots, C, \tag{4.2}$$
and form the output vector using $P_1,\ P_2,\ \cdots,\ P_C$.

After this stage, the local features have been transformed into M output vectors, each of dimension C, from the M local classifiers. The remaining question is how to use the local decisions efficiently to reach a final decision. In our scheme, by first concatenating the outputs of all local classifiers for one facial expression image, we generate its intermediate feature matrix of size $C \times M$. We then

apply PCA along with FLD analysis in order to project the intermediate feature matrices onto a discriminating, low-dimensional subspace, which facilitates an effective classification.

As discussed in Chapter 3, in order to extract desired number of discriminating components and avoid over-fitting, we adopt the recursive FLD (or RFLD) analysis [86] with regularization method. In this manner, we control the final performance of the proposed scheme by the two parameters, β (regularization factor) and N_c (number of discriminating components).

In the final decision-making stage, the global features after RFLD analysis are first normalized to have zero mean and unit standard deviation, and then fed into a nearest-neighbor classifier, which assigns the label of an expression to the input image.

4.2 Experimental Results

We have used the following facial expression databases: (1) Japanese Female Facial Expression (JAFFE) database; (2) Cohn-Kanade (CK) database [40]; (3) Taiwanese Facial Expression Image database (TFEID); (4) Yale-A database (YALE) [87]; and (5) The FG-NET Database with Facial Expressions and Emotions (FEED) [82].

The JAFFE database contains 213 images of 7 facial expressions of 10 Japanese female models, including 6 basic facial expressions (*happy, sad, angry, surprised, disgusted and scared*) and 1 neutral face. The CK database contains facial expression images of 100 university students (35 males and 65 females, 15% African-American and 3% Asian or Latino). Each person has a set of image sequences from neutral to certain facial displays coded by action units. For this study, we first select those sequences (from 94 persons) that can be labeled using one of the six basic expressions, and then for every person, we select 6 images for every expression (including neutral face). Note that, for some particular persons in the CK database, some of the expression images are missing. Therefore, we use 2581 images in the CK database for the experiments.

In addition, TFEID database contains 268 facial images with 7 expressions from 40 persons; and the YALE database, 75 facial images with 3 basic expressions

Figure 4.1: Sample images in the JAFFE database and the universal neutral face.

(happy, sad and surprise) and neutral face from 15 persons. Moreover, the FEED database contains a set of image sequences with 7 expressions from 18 individuals, from which we select 3516 images for this study.

4.2.1 Statistical Properties of COB Coefficients

Before using the COB coefficients to classify expressions, we conduct experiments using the JAFFE database to investigate the statistical properties of these coefficients. Figure 4.1 shows some samples of the JAFFE database and the synthesized frontal image described in [80] which is chosen as the universal neutral face in our experiments. Figure 4.2 shows some expressions and neutral flow-matrices as images. As discussed earlier, we observe that the images of expression flow-matrices still contain a significant amount of identity information.

In what follows, we represent all the images in the JAFFE database using the COB, and feed the COB coefficients to a self-organizing neural network [42] of size 30×30. See Figure 4.3 for the generated self-organizing map (SOM) in which labels $1-7$ stand for 6 basic expressions and the neutral face, respectively. Interestingly, the SOM exhibits distinct clusters for various expressions, thereby demonstrating that the COB coefficients have the property of homogeneity and, therefore, constitute meaningful features for expressions.

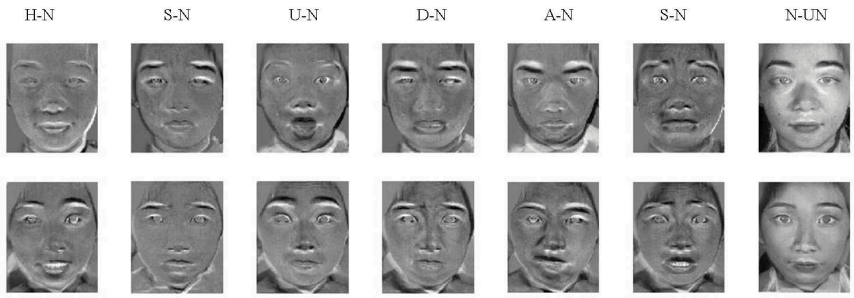


Figure 4.2: Flow-matrices as images for the JAFFE database. The left 6 columns contain expression flow-matrices of 6 basic expressions as images, whereas the last column contains neutral flow-matrices as images corresponding to different persons.

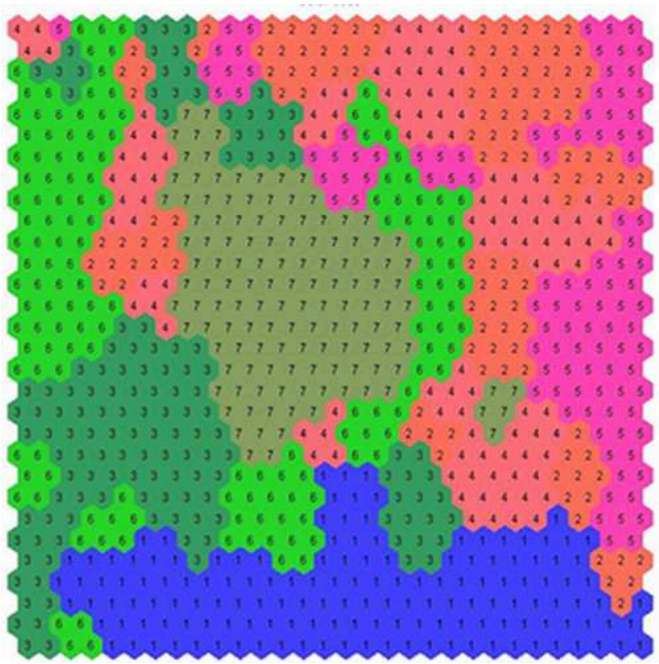


Figure 4.3: SOM of the COB coefficients obtained from the JAFFE database.

4.2.2 Cross Database Test Using COB with Local Methods

For the cross-database test, we use the following experimental scheme, and the universal neutral is the same as the one mentioned before:

- The combination of the CK and JAFFE databases is chosen as the training set, since the CK database has only a few images of Asians.
- The TFEID, YALE and FEED databases are chosen as the test set.

All the images are first preprocessed to have a uniform size of 150×120, and then divided into several local blocks. Every local block is represented using the COB, and the COB coefficients are fed to the local classifiers. Note that there are two key parameters, β and N_c, which control the final performance of the proposed FLD-based classifier combination. Therefore, we use the following validation strategy in order to determine the optimal values of β and N_c :

- Divide the training set (of the CK and JAFFE databases) into 10 groups according to identity information such that different groups contain images from different persons.
- Use one group as the validation set, and the other groups to train the system. Repeat until every group has been chosen as a validation set once. Average the recognition accuracy of validation sets.
- Use the simultaneous perturbation stochastic approximation (SPSA) algorithm [69], which is an effective optimization method, to find the optimal values of β and N_c for validation sets.

By using the validation set, we find that $\beta = 1.9$ and $N_c = 17$. We fix these parameters for the test sets (which are the TFEID, YALE and FEED databases). The results are given in Table 4.1, from which we find that 49 local blocks are the best trade-off between performance and effectiveness. Meanwhile, the recognition accuracies of the three test sets are all above 60%. Since the test images are from three totally different databases, we conclude that COB can generalize well to recognize facial expressions of strangers.

Table 4.1: Recognition results (%) of COB on cross databases with varying local blocks (LBs).

	9 LBs	25 LBs	49 LBs	81 LBs	121 LBs
FEED	59.77	64.61	**67.84**	65.22	62.02
YALE	60	58.33	**61.67**	58.33	61.67
TFEID	72.76	73.88	**73.88**	70.52	69.78

4.2.3 Individual Database Test Using COB with Local Features

Since most of the existing algorithms for facial expression recognition are applied to specific individual databases, i.e., the training and testing images belong to the same database, we investigate, for comparison, the performance of the COB with local features on the JAFFE and CK databases, separately. Since the JAFFE database contains limited samples, thereby leading to possible instability in the performance of an automatic recognition system, we adopt the "leave-one-person-out" cross-validation strategy. Further, since there are 100 subjects in the CK database, the above "leave-one-person-out" cross-validation is time-consuming. Alternatively, we randomly separate the subjects into $n = 10$ groups which are approximately equal in size, and adopt the "leave-one-group-out" cross-validation strategy. This procedure is similar to the method mentioned above. Since the same subjects will not appear in both training and testing, it is also person-independent. For the experimental setup, we preprocess the images as before, and set $\beta = 1.9$ and $N_c = 17$, and the number of local blocks = 49. The recognition rates for the JAFFE and CK databases are **84.38%** and **96.19%**, respectively.

4.3 Discussions

As applied to the JAFFE and CK databases for person-independent facial expression recognition, we first compare the accuracies of the combination of COB representation and local features with other results of the literature. From Table 4.2, we observe that COB gives the best result (**84.38%**) on the JAFFE database,

which is substantially higher than that of the rest. For the CK database, we observe that the recognition accuracy in [92] for 6 basic expressions is 96.26%. However, since the dynamic analysis method in [92] uses the neutral face as the basis to derive the motion of the facial expression sequence, the neutral face itself cannot be recognized at the same time. In contrast, COB is designed to be capable of recognizing not only the 6 basic expressions but also the neutral face. If we compare the present results with those for 7 classes which include the neutral face [3, 65], COB's recognition accuracy of **96.19%** turns out to be the highest. Therefore, in general, we conclude that, when applied to the CK database, the expression recognition accuracy of COB is quite comparable to that of the others in the literature.

Table 4.2: Comparison with Different Approaches on the JAFFE and CK Databases.

	[18]	[30]	[85]	[92]	[65]	[3]	COB
JAFFE	77%	79.21%	77.05%	-	-	-	84.38%
CK	-	-	-	96.26%	88.4%	86.9%	96.19%

4.4 Summary

In an attempt to simulate the expression-selective cells in STS, we have proposed a composite orthonormal basis (COB) to separate expression from identity information from face images. The combination of COB with local methods significantly improves the accuracy of facial expression recognition. Further, by conducting appropriate tests, we have also demonstrated, *for the first time*, that the proposed COB can be generalized to recognize expressions of faces of strangers from different databases.

Chapter 5

Facial Expression Recognition using Radial Encoding of Local Gabor Features and Classifier Synthesis

In the previous chapters, we have subsequently demonstrated that radial encoding, Gabor filters, local methods and COB algorithm, inspired by retinal ganglion cells, V1 cells, face-selective cells and expression-selective cells, can efficiently improve the performance of facial expression recognition. Here we will integrate these bio-inspired approaches with typical statistical approaches and propose a new hierarchical facial expression recognition scheme, aiming at building a universal expression recognizer that recognizes expressions from arbitrary given face images. In this scheme, based on the retinotopic mapping, the radial encoding strategy is extended for local Gabor outputs to obtain salient local features representing facial expressions. Then PCA and FLD analysis are applied to process the local features, the outputs of which are fed to local classifiers. The outputs from the latter, in turn, are then concatenated to form global, intermediate-level features which are subjected to the next level of PCA and FLD projections in order to extract the salient expression information, leading to classification by the global classifier.

5.1 General Structure of the Proposed Facial Expression Recognition Framework

The proposed facial expression recognition framework comprises four major steps as shown in Figure 5.1: (A) **preprocessing and partitioning**; (B) **local feature extraction and representation**; (C) **classifier synthesis** (to integrate local features); and (D) (final) **decision-making**. Below, we describe each of these steps.

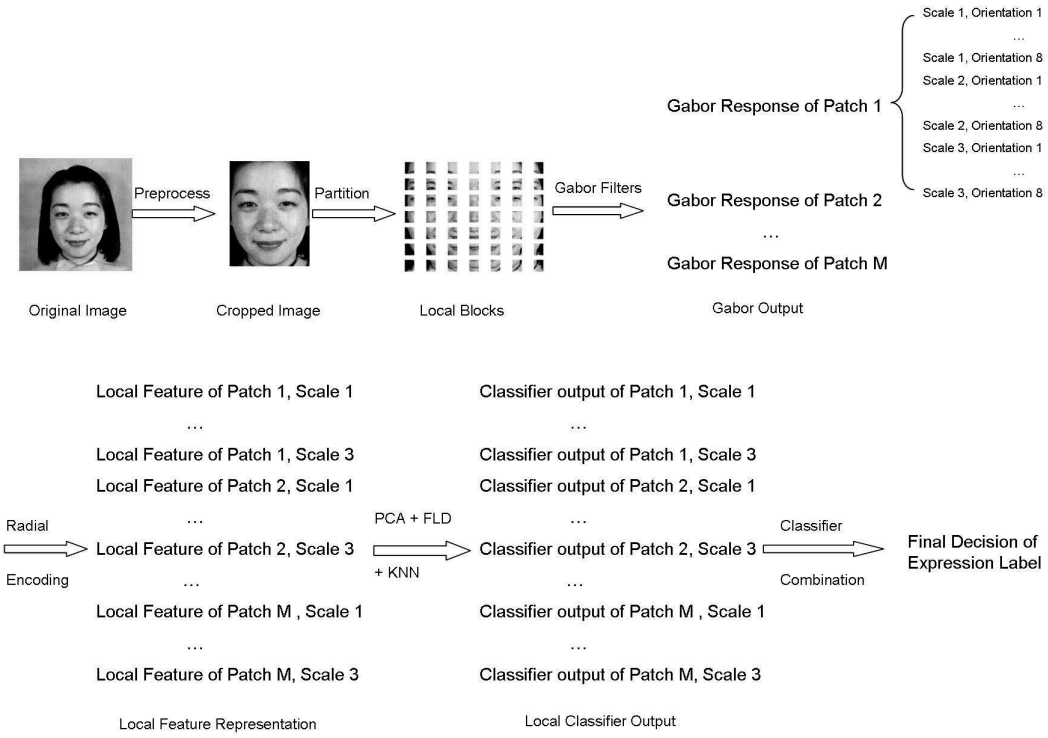


Figure 5.1: Flowchart of the *proposed* facial expression recognition framework.

5.1.1 Preprocessing and Partitioning

With a view to confine the processing only to the required parts of an image, we first manually determine eyes' and mouth's positions of each image and use these positions to obtain facial region of each image, simultaneously achieving a uniform image size of 184×152. Then, we divide each image into several local regions,

some of which contain certain facial components (Figure 5.1), like the *eyebrow corner, mouth corner and wrinkle*, which are critical for recognizing expressions. We employ these regions as the local receptive fields of the basic units which extract low-level local features, imitating the V1 cells in human vision system. It has been found, in neurophysiological and psychological studies [33, 64], that two neighboring cells (both in retina and visual cortex) usually have overlapping receptive fields (see Figure 5.2 below). Therefore, in our implementation, the neighboring local regions are designed to have 50% overlap. The number of the local regions is determined by the ratio (ρ) of size of the input image to that of the local region. Assuming that the ratio ρ is the same for the height and the width of the image, the total number of local blocks is $(2\rho - 1)^2$ due to the (assumed) 50% overlap.

The actual number of regions is chosen on the basis of the following considerations: 1) the facial components should be as complete as possible in the local regions; and 2) the local regions should be small enough such that local features could be extracted from the facial components. For comparison, Figure 5.2 shows local regions generated by three ratios: 3, 4, and 5, resulting in 25, 49, and 81 local regions respectively. The satisfactory number of local blocks needed is determined using trial and error, as a trade-off between recognition accuracy and computational load (see Section 5.2.2 below for details).

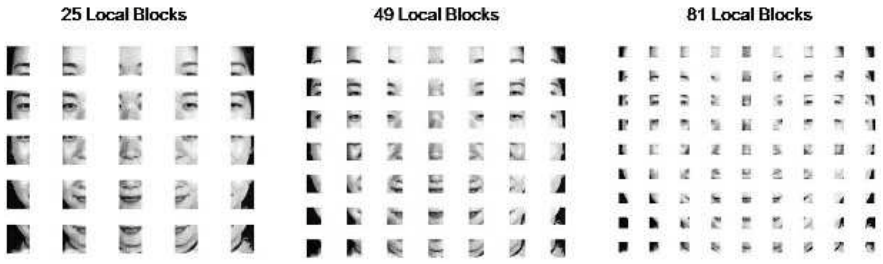


Figure 5.2: Local blocks with different sizes.

5.1.2 Local Feature Extraction and Representation

Each region R of the input image I is subjected to a set of Gabor filters. Recall that Gabor filters can be described by:

$$G(x,y) = \exp(-\frac{\widehat{x}^2 + \gamma^2\widehat{y}^2}{2\sigma^2}) \times \cos(\frac{2\pi}{\lambda}\widehat{x} + \varphi), \tag{5.1}$$

where

$$\widehat{x} = x\cos\theta + y\sin\theta, \quad \widehat{y} = -x\sin\theta + y\cos\theta, \tag{5.2}$$

and (x, y) refers to the pixel position in a $2D$ coordinate system, and the parameters affecting the filter outputs are: θ (orientation), γ (aspect ratio), σ (effective width), φ (phase), and λ (wavelength). These parameters can be chosen such that the filters model the tuning properties of cortical cells [64].

We consider (i) eight orientations (from 0 to $7/8\pi$ in uniform steps of $1/8\pi$) in order to capture subtle changes of facial components, and (ii) three filter sizes ($\mathcal{S}$ = 11, 20 and 29) in order to determine σ and λ according to the following formulas [64]:

$$\sigma = 0.0036\ \mathcal{S}^2 + 0.35\ \mathcal{S} + 0.18, \quad \text{and} \quad \lambda = \frac{\sigma}{0.8}. \tag{5.3}$$

The remaining two parameters, φ, and γ, are fixed to 0 and 0.5, respectively, since they are found to have little influence on filter's tuning properties [64]. Therefore, for each local region of an input image, we obtain $3 \times 8 = 24$ Gabor filter outputs which we need to efficiently encode to form feature-arrays for further processing.

Figure 5.3 shows the mapping of the retina (A) on the lateral geniculate nucleus (B) and the primary cortex (C) in the macaque monkey [9]. We observe that for a visual stimulus formed on the retina, the neighboring retinal regions are represented by neighboring regions of the visual cortex. Moreover, this retinotopic representation in the cortical areas is nonlinear, due to the fact that the retinal fovea is disproportionately mapped in a much larger area of the primary cortex than the retinal periphery. Motivated by this evidence of the retinotopic mapping in the visual cortex [75], we encode the Gabor filtered outputs (of all the local regions obtained from a facial image) using a radial grid structure, imitating that of the retina itself. The grid resolution is chosen such that the inner sector captures as small a number of pixels as possible. Note that the area of the inner sector is much smaller than that of the outer sector. Therefore, the inner sector contains less pixels than the outer sector. A consequence is that an averaging operation in a sector in the center of the radial grid (which has fine resolution)

downsamples the input data in contrast with a similar operation in the periphery of the radial grid which has a coarse resolution. In our implementation, a grid of size 18×5 (angular *vs* radial) is chosen such that the innermost sector contains at least one pixel, thereby resulting in a feature matrix of size, 18×5, for each Gabor filtered local block. As demonstrated in [28], a radial grid with too low a resolution fails to capture discriminative features, while a radial grid with too high a resolution increases computational load without a significant improvement in the final result. Here the size 18×5 has been chosen on the basis of the experimental results and with the aim of imitating retinotopic mapping.

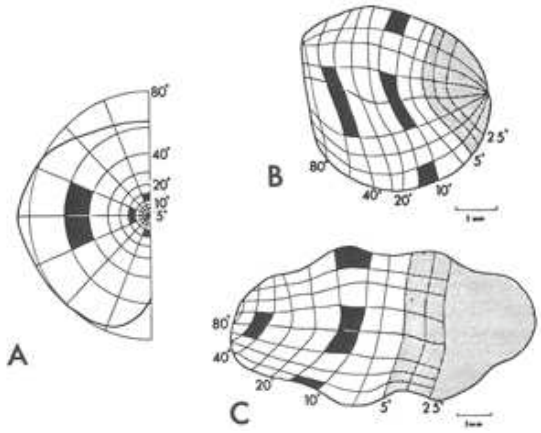


Figure 5.3: Retinotopic mapping from retina to primary cortex in the macaque monkey.

See Figure 5.4 which shows an example of the radial grid placed on a part of the facial image. The actual radial encoding procedure on Gabor outputs is as follows:

1. Place a radial grid of resolution 18×5 on a Gabor filtered image, with the center of the radial grid at the center of the local region, and with radius r of the outermost circle of the radial grid determined by $r = min(w, h)/2$, where w and h are the width and height of the local region, respectively.

2. For each sector in the grid, with radial grid coordinates (i, j), where $i = 1, 2, \cdots, 18$ and $j = 1, 2, \cdots, 5$, compute $v(i, j) = P\_sum\_value/P\_num$, where P_num is the number of pixels that fall into that sector, and P_sum_value,

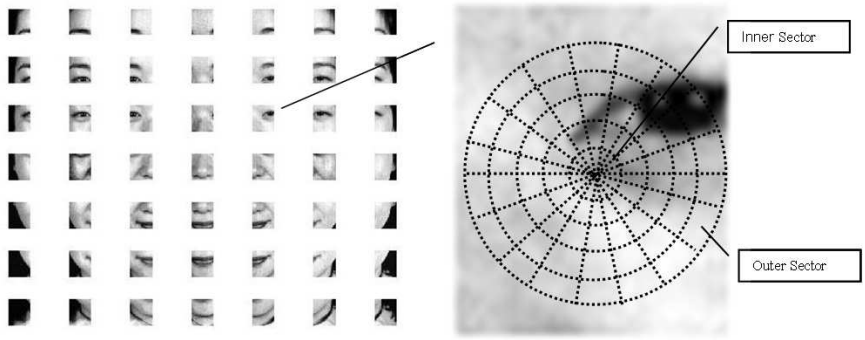


Figure 5.4: Example of the radial grid placed on a gray-level image.

the sum of all the pixel values in that sector. Essentially, $v(i, j)$ represents the averaged pixel value for sector (i,j).

3. Form the Gabor feature matrix, $\{v(i, j)\}$, where $i = 1, 2, \cdots, 18$ and $j = 1, 2, \cdots, 5$.

After the radial encoding of each local region in the feature space, the 24 Gabor filtered outputs are represented by 24 feature matrices of size 18×5. Then, we group the feature matrices obtained from Gabor filtered images with the same scale but different orientations together, resulting in 3 new feature matrices of size 90×8 for each local region, where 90 is the number of cells $= 18 \times 5$ in the radial grid, and 8 is the total number of orientations used in Gabor filter. The 3 local feature representations corresponding to 3 different Gabor filter scales are obtained in every local region. In general, a facial image is represented by 147 (i.e., 49 local regions at 3 scales) local features, each of which is a matrix of size 90×8.

A straightforward and common way to build the global feature is to group the 147 local features together to form a new feature vector for every image. However, we discover that this kind of a direct global Gabor feature is related to identity rather than expression. In fact, when we apply the ISODATA clustering algorithm [35] to the global Gabor feature, the results (see Section 5.2.1) clearly show that its direct use is not suitable for recognizing expressions. Recall that recent biological findings [76, 77, 78] on face processing in the human brain suggest the following:

1. Face detection and its simultaneous identification, and further processing for its expression recognition.

2. Capturing local facial information in each cell acting as a local receptive field.

3. Possible reconstruction of a face, preserving most facial information, by combining local information.

The above empirical findings justify the further processing of the local Gabor features in order to build global features related to expressions.

5.1.3 Classifier Synthesis

The local features are now to be integrated into intermediate-level features for a global representation of facial expressions. In the absence of any tangible evidence for any explicit mechanism of a biological feature combination in the human brain, we resort to a statistical method to integrate local features. More specifically, for each local feature, a local classifier makes a local decision, and the outputs of all the local classifiers are then used to generate an "accumulated" decision.

Before feeding local features into local classifiers, FLD analysis is used to seek a projection for each local feature such that it can be optimally separated from the others. Moreover, following the procedure suggested in [4], before invoking FLD analysis, we first use PCA to reduce the dimension of the local features, and select $(n - C)$ principal components to represent the input data so that the matrix S_w becomes nonsingular for FLD analysis. Here, n is the sample size; C is the number of classes; and S_w is the within-classes scatter-matrix in the FLD calculation.

In our implementation, another problem caused by PCA and FLD analysis is over-fitting. That is, due to the small sample size of training set, most of the feature vectors are mapped to one specific point in the feature space after training. In order to overcome this problem, a regularization method can be used [29]. To this end, we have explored the addition of $Ave\ (Eigv(S_W)) \cdot I$ to S_W in FLD, where $Ave\ (Eigv(S_W))$ denotes the average eigenvalue of S_W, and I is the identity matrix. The experimental results show that such a term can lead to

a satisfactory performance which is close to the optimal value (see Section 5.2.2 for details). The resulting $(C-1)$-dimensional vectors are normalized to have zero mean and unit standard deviation, before using them as the input to a local classifier for each local feature.

The local classifier we choose is the modified k-nearest neighbor (KNN) with $k = 1$:

1. For a given test input vector, compute the Euclidean distances from the test input to all training vectors;
2. For each class, find the minimum distances d_i, where $i = 1, 2, \cdots, C$;
3. Estimate the probabilities for each class using the following indication function:
$$P_i = \frac{(1+d_i)^{-1}}{\sum_{k=1}^{C}(1+d_k)^{-1}}, i = 1, 2, \cdots, C, \tag{5.4}$$
and form the output vector using $P_1,\ P_2,\ \cdots,\ P_C$.

Note that the choice of the indication function affects the final result as discussed in [62]. However, the Euclidean distance-based indication function has been found to lead to the best performance not only in our implementation but also in the literature [62].

After this stage, the 147 (49×3) local features have been transformed into 147 C-dimensional output vectors from the 147 local classifiers. The remaining question is how to use the local decisions efficiently to reach a final decision. In our scheme, by first concatenating the outputs of all local classifiers for one facial expression image, we generate its intermediate feature matrix of size $C \times 147$. We then apply PCA along with FLD analysis in order to project the intermediate feature matrices onto a discriminating, low-dimensional subspace, which facilitates an effective classification.

Recall that since the FLD can at most extract $C-1$ discriminating components from the input data, which may be insufficient to represent the global features with high complexity, the recursive FLD (RFLD) analysis is also adopted. Further, in order to avoid over-fitting, we invoke the regularization method in RFLD analysis as follows:

$$S_W \rightarrow S_W + \beta \cdot Ave(Eigv(S_W)) \cdot I, \tag{5.5}$$

where β is the regularization factor which controls the influence of the regularization term (note that the other terms, $Ave(Eigv(S_W))$ and I have been defined earlier). In this manner, we control the final performance of the proposed scheme by the two parameters, β and N_c (which is the number of extracted features), whose effect on the final recognition performance is studied in Section 5.2.2.

5.1.4 Final Decision-Making

In the final decision-making stage, the global features after RFLD are first normalized to have zero mean and unit standard deviation, and then fed into a nearest-neighbor classifier, which assigns the label of an expression to the input image.

5.2 Experimental Results

The demo of the proposed framework is implemented in MATLAB™, and all simulations are conducted using a 2.66 GHz Intel® Core™ Quad processor with 8 GB memory. We use the following two facial expression databases for experiments: (1) Japanese Female Facial Expression (JAFFE) database; (2) Cohn-Kanade (CK) database.

5.2.1 ISODATA results on Direct Global Gabor Features

First of all, we present experimental results of using ISODATA to cluster the direct global Gabor features discussed in Section 5.1.2. The database used here is JAFFE. We denote the 10 different expressers by P_i, where $i = 1, 2, \cdots, 10$; and the 7 different expressions by E_i, where $i = 1, 2, \cdots, 7$. The direct global Gabor features of all the images are fed into ISODATA algorithm. We conduct the following two experiments in order to verify whether the global Gabor features characterize identity or expression or both:

1. Investigate the homogeneity of the direct global Gabor features with respect to identity. By setting the number of desired clusters to be 10, after clustering using ISODATA, check the identity distribution within every cluster.

2. Investigate the homogeneity of the direct global Gabor features with respect to expression. By setting the number of desired clusters to be 7, after clustering using ISODATA, check the expression distribution within every cluster.

Tables 5.1 and 5.2 show the results of the above two experiments respectively. The actual resulting number of clusters, which is determined by the ISODATA algorithm automatically, might not necessarily be the same as the desired number of clusters set by the user at the beginning. Note that although there are 11 clusters in Table 5.1, cluster 4 and 9 correspond to the same person, P9. From Table 5.1, we observe that the feature data group into clusters according to the identity information, while, from Table 5.2, we observe that feature data representing different expressions are mixed in every cluster, which implies that the resulting 7 clusters do not correspond to the 7 expressions at all. These results confirm that the global Gabor features are more characteristic of identity than of expression.

Table 5.1: ISODATA results on direct global Gabor features with respect to identity.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Cluster 1	0	0	0	0	0	0	0	0	0	**21**
Cluster 2	0	0	0	0	**20**	0	0	0	0	0
Cluster 3	0	0	0	**22**	0	0	0	0	0	0
Cluster 4	0	0	0	0	0	0	0	0	**9**	0
Cluster 5	0	0	**22**	0	0	0	0	0	0	0
Cluster 6	0	0	0	0	0	0	0	**19**	0	0
Cluster 7	0	0	0	0	0	0	**19**	0	0	0
Cluster 8	**22**	0	0	0	0	0	0	0	0	0
Cluster 9	0	0	0	0	0	0	0	0	**12**	0
Cluster 10	0	**1**	0	0	0	**14**	**2**	0	0	0
Cluster 11	0	**22**	0	0	0	**7**	0	**1**	0	0

Table 5.2: ISODATA results on direct global Gabor features with respect to expression.

	E1	E2	E3	E4	E5	E6	E7
Cluster 1	6	5	6	5	6	6	6
Cluster 2	3	0	0	0	0	3	3
Cluster 3	1	1	10	0	0	3	3
Cluster 4	0	3	3	5	3	0	0
Cluster 5	4	3	3	3	2	4	3
Cluster 6	10	12	3	9	9	10	9
Cluster 7	8	6	5	7	10	6	6

5.2.2 Experiments on an Individual Database

We first study the performance of the proposed framework on a specific database. That is, the training images and testing images are from the same database. As discussed in Chapter 4, the person-independent cross-validation strategy is adopted for the individual database test.

5.2.2.1 Effect of Number of Local Blocks

Table 5.3 shows the recognition results for different numbers of local regions on the JAFFE and CK databases. The other parameters, β and N_c in the final stage, are chosen to be the default values (i.e., $\beta = 1$ and $N_c = C - 1$, where $C = 7$ is the number of expression classes under study). We discover that 81 local blocks lead to a recognition accuracy comparable to or even higher than the accuracy from 49 local blocks, but at a greater computational cost. As already discussed in Section 5.1, we conclude that the choice of 49 local blocks is reasonable, giving high accuracies on the two databases. Therefore, for all the following experiments, we assume that the input image is divided into 49 local blocks.

5.2.2.2 Effect of Radial Grid Encoding on Gabor Filters

Table. 5.4 shows the effectiveness of radial encoding for downsampling the Gabor filtered outputs. Here the experiments are conducted only on the JAFFE

Table 5.3: Recognition rates (%) on JAFFE and CK for different NO. of local blocks.

	25 LBs	49 LBs	81 LBs
JAFFE	84.03	87.32	87.32
CK	84.75	89.07	88.80

database. In the final stage of the proposed classifier combination method, $\beta = 1$, and $N_c = 7$. We also list the results of using the Borda count and decision template as the combination method in the final stage for comparison. For extracting local features, we consider the following 5 different methods in the local feature extraction and representation module in our system:

- Gray-level images alone;
- Radial encoding on gray-level images;
- Gabor filters on gray-level images;
- Gabor jets generated by uniformly downsampling Gabor filtered outputs (based on gray-level images); and
- Radial encoding on Gabor filtered outputs (based on gray-level images).

We keep the other setups the same for these 5 cases in order to conduct a fair comparison. Experimental results show that the proposed radial grid encoding can efficiently downsample the Gabor filtered outputs and also significantly improve the final recognition accuracy. For a further comparison of the proposed algorithm with different approaches, see Section 5.3.

5.2.2.3 Effects of Regularization Factor and Number of Components

Figure 5.5 shows the recognition results of the proposed framework for different values of regularization factor β in the final stage with three different values of Nc (6, 15 and 25). It is observed that the regularization factor, β, significantly influences the final recognition result, and the choice of $\beta = 1$ leads to a recognition accuracy close to the optimal value. However, the optimal value of β is

Table 5.4: Recognition rates (%) on JAFFE with different local feature encoding methods.

	Decision Template	Borda Count	Proposed Method
Case 1	65.26	69.48	69.01
Case 2	67.61	70.42	70.89
Case 3	76.06	75.59	77.00
Case 4	79.81	83.10	82.16
Case 5	**82.63**	**87.32**	**87.32**

problem-dependent, and is affected by the number of components N_c, which by itself is another parameter for tuning the RFLD in the final stage. Therefore parameter tuning techniques can be utilized to achieve a satisfactory performance. However, since most of the algorithms in the literature do not benefit from this kind of parameter tuning, we adopt the following strategy to determine β and N_c for a fair comparison.

1. Use the TFEID database as the data set for parameter tuning;

2. Divide the TFEID database into 10 groups according to the identity information. And use one group as the validation set, and the other groups to train the system. Repeat until every group has been chosen as a validation set once. Average the recognition accuracy of validation sets;

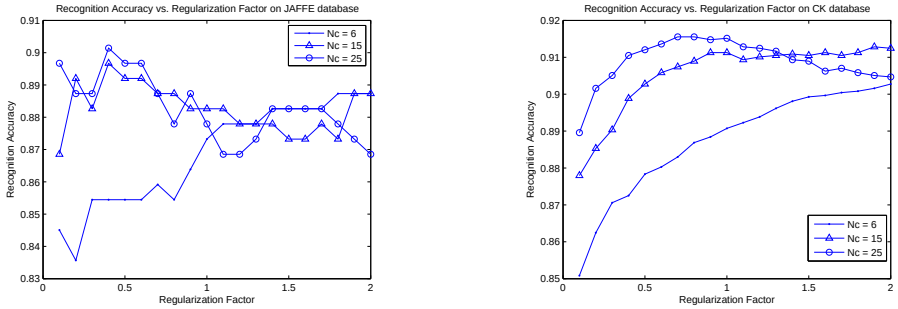
3. Use the SPSA algorithm to find the optimal values of β and N_c for validation sets. In an attempt to reduce the computational load, we search for (i) β in $[0, 3]$ in steps of 0.1, and (ii) N_c in $[6, 36]$ in steps of 1 within 100 iterations.

Based on the TFEID database, we find the optimal values for β and N_c are 1.1 and 17, respectively. We fix these values for JAFFE and CK database in the individual database test. Tables 5.5 shows the highest recognition results of our system with specified β and N_c. On comparison with the recognition results obtained from using the Borda count and decision templates in the final stage, we find that the proposed classifier combination outperforms both.

Tables 5.6 and 5.7 show the confusion matrix of the best result of the proposed framework on JAFFE and CK databases, respectively. We observe the following:

1. For the JAFFE database, happy, sad, angry, and surprise expressions are easy to recognize, while disgusted, neutral and scared expressions are not. Also, the recognition accuracy for the scared expression is the lowest. This is consistent with the fact that expressers of JAFFE database find it most difficult to deliver the scared expression accurately, as reported in [91].

2. For the CK database, happy, surprise, neutral and disgusted expressions are easy to recognize while angry, sad and scared expressions are not.



(a) Recognition rates on JAFFE (b) Recognition rates on CK

Figure 5.5: Recognition rates with different regularization factors and number of discriminating features.

Table 5.5: Highest recognition results (%) of our system on the JAFFE and CK databases.

	Decision Template	Borda Count	Proposed Method
JAFFE	82.63	87.32	**89.67** ($\beta = 1.1, N_c = 17$)
CK	63.58	67.30	**91.51** ($\beta = 1.1, N_c = 17$)

5.2.3 Experiments on Robustness Test

In this section, we consider the following issue: *Is the proposed framework for recognizing facial expressions robust to corrupted data and to missing information?* It is interesting that, in [43], facial expression recognition under occlusions (i.e., missing information) is cited as a related challenge.

Table 5.6: Confusion Matrix (%) for the best result of our system on the JAFFE database.

	Happy	Sad	Surprise	Disgusted	Angry	Scared	Neutral
Happy	**100**	0	0	0	0	0	0
Sad	0	**93.34**	0	0	3.33	3.33	0
Surprise	0	0	**96.67**	0	0	3.33	0
Disgusted	0	6.9	0	**86.21**	3.45	3.44	0
Angry	0	0	0	6.67	**93.33**	0	0
Scared	3.12	9.38	3.12	9.38	0	**75.00**	0
Neutral	3.33	3.33	0	0	6.67	0	**86.67**

Table 5.7: Confusion Matrix (%) for the best result of our system on the CK database.

	Happy	Sad	Surprise	Disgusted	Angry	Scared	Neutral
Happy	**93.33**	0	0	0	0	1.57	5.10
Sad	0	**90.03**	0	0	4.16	0	5.81.
Surprise	0	0	**97.67**	0	0	0	2.33
Disgusted	0.45	0	0	**95.93**	1.36	0	2.26
Angry	0	0.44	0	9.69	**75.33**	2.21	12.33
Scared	5.53	0	2.08	2.08	1.73	**85.47**	3.11
Neutral	0.37	0.74	0.92	0.74	2.93	0.18	**94.12**

In order to answer the above question, we once again use the CK database but with one of the eyes and/or mouth masked, thereby simulating occluded facial images. Figure 5.6 shows some samples of such images with different sizes of masks. Next, we crop the input images in order to have a uniform size (184×152) in the pre-processing stage, and then divide them into 49 local blocks. The two parameters β and N_c are set to 1.2 and 18, respectively, as explained in Section 5.2.2.3 above. See Table 5.8 for the classification results which correspond to person-independent cross-validation strategy.

In contrast, in [43], the authors use another kind of cross-validation strategy: randomly dividing the CK database into 5 groups, and choosing one group as a test set while choosing the remaining as the training set, and repeating the experiment until all the groups have been used once as a test set.

For comparison, we also conduct experiments using the same random cross-

(a) Eye masked with small occlusion.

(b) Mouth masked with small occlusion

(c) Eye masked with medium occlusion.

(d) Mouth masked with medium occlusion

(e) Eye masked with large occlusion.

(f) Mouth masked with large occlusion

Figure 5.6: Masked samples in the CK database.

Table 5.8: Recognition rates (%) on the masked CK using person-independent cross-validation.

	small mask	medium mask	large mask
Eye	91.32	90.47	88.57
Mouth	85.97	82.91	76.32
Normal		91.51	

validation strategy. For the results, see Table 5.9 where the two parameters β and N_c are set to 1.5 and 23, respectively. The parameters are determined by applying the SPSA algorithm to the normal case (without occlusions) using the random cross-validation strategy.

From Tables 5.8 and 5.9, we observe that expressions with a masked mouth are more difficult to recognize than those with masked eyes. This is, in fact, consistent with human perception. For example, if we examine Figure 5.6 (f), we observe that the sad and neutral expressions are almost the same when the mouth is masked. However, from Figure 5.6 (e), we observe that the sad and neutral expressions are still distinctive when the eyes are masked. Furthermore, for *sad* and *scared* expressions, the discriminating information mainly distributes in the lower part of faces, thereby substantially reducing the performance of recognizing *sad* and *scared* expressions with large mouth masks, as shown in Table 5.10.

Table 5.9: Recognition rates (%) on the masked CK database using random cross-validation.

	small mask	medium mask	large mask
Eye	99.46	99.23	98.88
Mouth	98.72	98.26	96.51
Normal		99.65	

Similarly, the discriminating information of *angry* expression mainly distributes in the upper part of faces, leading to substantially lower recognition accuracies of *angry* expression, as shown in Table 5.10 and 5.11.

Table 5.10: Confusion Matrix (%) using person-independent cross-validation on the CK database with large mouth masks.

	Happy	Sad	Surprise	Disgusted	Angry	Scared	Neutral
Happy	**80.20**	0.39	1.18	0	0	11.18	7.05
Sad	0	**50.97**	9.70	0	10.25	1.66	27.42
Surprise	0	0.23	**93.01**	0	0.93	0.47	5.36
Disgusted	4.52	0.45	0	**85.53**	6.33	0	3.17
Angry	0.88	11.45	2.65	10.57	**62.12**	1.76	10.57
Scared	3.46	1.73	11.76	3.11	3.46	**69.55**	6.93
Neutral	0.74	6.25	3.31	2.02	4.77	0.74	**82.17**

Table 5.11: Confusion Matrix (%) using person-independent cross-validation on the CK database with large eye masks.

	Happy	Sad	Surprise	Disgusted	Angry	Scared	Neutral
Happy	**89.80**	0	0	0.59	0.59	4.71	4.31
Sad	0	**86.43**	0	0	2.77	0	10.80
Surprise	0	0	**96.97**	0	0	0	3.03
Disgusted	1.36	0	3.17	**91.86**	0.90	0	2.71
Angry	0	6.17	0	9.25	**63.00**	1.76	19.82
Scared	7.27	0	0	3.11	2.08	**84.78**	2.77
Neutral	0.92	1.10	0	1.29	2.94	0.18	**93.57**

Moreover, in [43], the best accuracy of recognition is 96.3% on eye occlusion with large masks, and 93.7% on mouth occlusion with large masks. Note that

these results have been obtained on the whole CK database which is slightly different from our experimental setup. Therefore, we can conclude that the proposed framework can also produce comparable performance on facial expression recognition under occlusions, thereby demonstrating its robustness.

5.2.4 Experiments on Cross Databases

In this section, we consider the following cross-database expression recognition problem: *Are the expression features, extracted by a recognition algorithm, representative enough such that a new facial expression image from another database can also be recognized?* It is well known that humans can recognize expressions of an unfamiliar person; that is, a change of identity does not seem to affect the representation of an expression in the human brain while recognizing it. However, for automatic expression recognition, it has turned out to be difficult to separate expression from identity. Moreover, images from different databases in general exhibit different illumination conditions or background information. Therefore, it appears that a recognition accuracy of even 50% can be considered acceptable for automatic expression recognition in a cross database problem.

With the above caveat in mind, we now conduct two cross database experiments: 1) the training database is JAFFE, but test database is CK; and 2) vice versa. Note that the JAFFE database contains sufficient variations of expressions, but only 213 images. Therefore, the final performance is probably limited by the small sample size if the JAFFE database is used for training. On the other hand, the CK database contains sufficient facial images from 100 different subjects, but only 3% of them are Asian or Latino. So, in this case, the final performance is probably affected by the differences of identity as well as of race.

Table 5.12 shows the corresponding experimental results of the proposed framework. Note that all the input images are first cropped to have a uniform size (184×152) in the preprocessing stage, and then are divided into 49 local blocks. We keep the parameters in the local feature extraction and integration stage the same as before, while in the final stage we tune the parameters β and N_c in order to obtain the optimal performance.

From Table 5.12, we find that:

- when we use the JAFFE database for training and the CK database for testing, the best recognition accuracy is 54.05%; but
- when we use the CK database for training and the JAFFE database for testing, the best recognition accuracy is 55.87%.

In the two cross database experiments mentioned above, both the recognition rates are higher than 50%, showing that the proposed framework with the proposed classifier combination achieves an acceptable performance on a cross-database test.

Table 5.12: Highest recognition results (%) of the proposed framework on the JAFFE and CK databases.

	Decision Template	Borda Count	Proposed Method
JAFFE Train, CK Test	34.41	52.11	**54.05** ($\beta = 1.2, N_c = 28$)
CK Train, JAFFE Test	27.23	32.86	**55.87** ($\beta = 2.5, N_c = 27$)

5.2.5 Experiments for Generalization Test

We extend the cross-database test in order to check the generalizability of the proposed framework. The experimental scheme is as follows: :

- The combination of the CK and JAFFE databases is chosen as the training set, since the CK database has only a few images of Asian.
- The databases TFEID, Yale-A (YALE) and FEED are chosen as test sets.

Recall that i) the TFIED database contains 268 facial images with 6 basic expressions and neutral face from 40 persons; ii) the YALE database contains 75 facial images with 3 basic expressions (happy, sad and surprise) and neutral face from 15 persons; and iii) the FEED database contains a set of image sequences with 6 basic expressions and neutral face from 18 individuals, and we select 3516 images from FEED as test samples. As before, all the input images are first cropped to have a uniform size (184×152) in the preprocessing stage, and then divided into 49 local blocks.

For the generalization test, we use the following strategy to determine the two key parameters β and N_c:

- Divide the training set (of the CK and JAFFE databases) into 10 groups according to the identity information.
- Use one group as the validation set, and the other groups for training. Repeat until every group has been chosen as a validation set once. Average the recognition accuracy of validation sets.
- Use the SPSA algorithm to find the optimal values of β and N_c for validation sets.

By using the validation set, we determine β and N_c to be 1.1 and 18, respectively. And we fix these parameters for the test sets (which are the TFEID, YALE and FEED databases). The results are given in Table 5.13, from which we observe that the recognition accuracies of three test sets are all above 60%. Since the test images are from the three totally different databases, we conclude that the proposed framework can generalize well for facial expression recognition from novel faces.

Table 5.13: Highest recognition results (%) of the proposed framework on the generalization test.

	TFEID	YALE	FEED
Proposed Method	61.94	60.66	61.43

5.3 Discussions

In this section, we first compare the proposed expression classification scheme with that of others as applied to the JAFFE and CK databases, using the person-independent strategy for cross-validation. Note that the results of different algorithms may not be directly comparable because of differences in experimental setups, the number of subjects and of expressions used, and so on, but they can still indicate the discriminative performance of every approach. Furthermore, it

should be added here that we present the best results on the individual database with the tuned values of the parameters β and N_c.

Table 5.14 shows the comparison with a few other approaches applied to the JAFFE database in the literature. The first 6 approaches belong to the global method, while the last 3 approaches, including ours, belong to the local methods. We observe that our approach, with radial encoded Gabor features, gives the best result (**89.67%**), on the JAFFE database, which is substantially higher than the rest.

Table 5.15 shows a similar comparison with respect to the CK database. The first 4 methods belong to dynamic analysis method which is applied to video sequences, while the last 5 approaches, including ours, belong to static analysis method which is normally applied to static images. We observe that the best recognition accuracy of 6 expressions on the CK database is 96.33%, reported in [47]. However, since this kind of dynamic analysis method uses the neutral face as the basis to derive the motion of the facial expression sequence, the neutral face itself cannot be recognized at the same time. In contrast, the proposed approach is designed to be capable of recognizing not only the 6 basic expressions but also the neutral face. If we compare our results to those for 7 classes which include the neutral face ([65] and [3]), our recognition accuracy of 91.82% is actually the highest. Therefore, in general, we can conclude that when applied to the CK database, the result of our approach (**91.51%**) is quite comparable to that of the others in the literature.

Table 5.14: Comparison with different approaches on the JAFFE Database.

	Subjects	Images	Classes	Method	Recognition Rate(%)
[52]	9	193	7	Gabor jets	75
[66]	9	193	7	Fisher weight maps	69.4
[85]	10	183	6	KCCA	77.05
[25]	10	183	6	Statistical features	62.78
[89]	10	213	7	Modified KCCA	67
[44]	10	213	7	Salient feature vectors	85.92
[18]	9	193	7	LBP	77
[30]	10	213	7	Enhanced LBP	79.21
Ours	10	213	7	Radial encoded Gabor jets	**89.67**

Table 5.15: Comparison with different approaches on the CK Database.

	Subjects	Classes	Dynamic	Measure	Recognition Rate (%)
[88]	97	6	Y	5-fold	90.9
[2]	90	6	Y	-	93.66
[92]	97	6	Y	10-fold	96.26
[47]	90	6	Y	leave-one-subject-out	96.33
[65]	96	7(6)	N	10-fold	88.4 (92.1)
[3]	90	7	N	10-fold	86.9
[50]	90	6	N	leave-one-subject-out	93.8
[74]	97	6	N	-	93.8
Ours	94	7	N	10-fold	**91.51**

5.4 Summary

In this chapter, we have proposed a hierarchical facial expression recognition framework in the form of a novel fusion of statistical techniques and the known model of a human visual system. An important component of this framework is the biologically-inspired radial grid encoding strategy which is shown to effectively downsample the outputs of a set of local Gabor filters as applied to local patches of input images. Local classifiers are then employed to make the local decisions, which are integrated to form intermediate features for representing facial expressions globally. The recognition accuracies obtained on application to standard individual databases have been shown to be significantly better than those in the literature. Furthermore, it is believed that our demonstration of a satisfactory cross-database recognition performance is the first of its kind. We have also demonstrated, by conducting appropriate tests, that the proposed system is robust to corrupted data and to missing information and can be generalized to recognize expressions of novel faces from wide-ranging databases.

Chapter 6

The Integration of the Local Gabor Feature Based Facial Expression Recognition System

After designing the framework of local Gabor feature based facial expression recognition in the previous chapter, we now consider the implementation of a practical application based on the proposed framework. Therefore, in this chapter, we focus on the practical issues of developing a web-based facial expression recognition system.

6.1 The Structure of the Facial Expression Recognition System

The flowchart of the system is shown in Figure 6.1. The steps involved in recognizing an expression from an image are briefly as follows.

Step 1: The face image to be analyzed is uploaded to the system to automatically check for a human face in it. If the uploaded image does not contain a human face, the system outputs an alert message asking another face image to be uploaded.

Step 2: The system detects facial components, and, for face normalization, computes the coordinates of and marks the centers of eyes and mouth. The

displayed marks are verified by the user for intervention and manual marking for better accuracy, if necessary. The system then normalizes the image such that the centers of the eyes and mouth of the input image match with those of the training images.

Step 3: The system crops the face region of the input image to remove background information.

Step 4: The system subjects the normalized image to the process of expression classification.

More details of the components of the system are presented below.

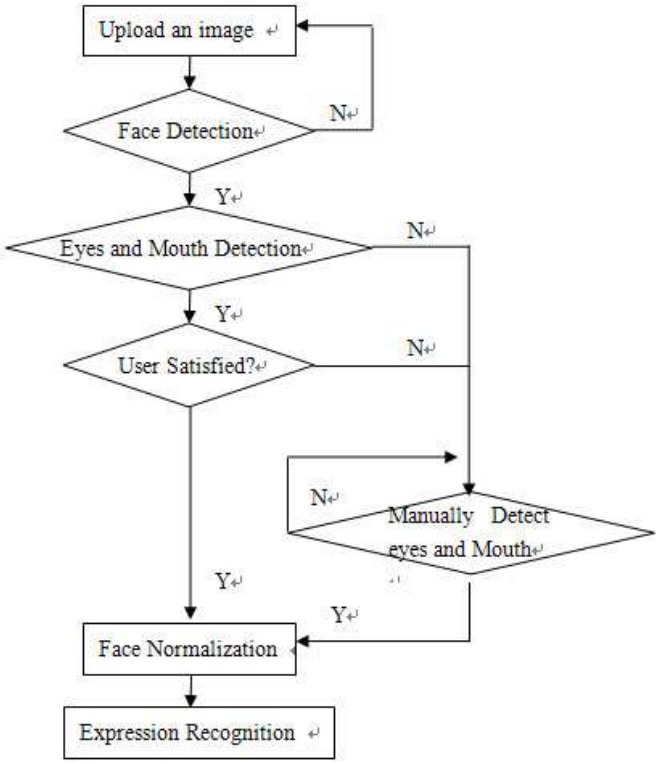


Figure 6.1: The flowchart of the proposed system.

6.2 Automatic Detection of Face and its Components

An image presented to the system is first verified for the existence of a human face in it. The real-time face detection algorithm of [81], which can be trained to detect a variety of object classes, has been implemented, and adopted in OpenCV [6]. The three major components of the face detection algorithm are:

1. Haar-like feature computation. As shown in Figure 6.2 , three kinds of Haar-like features are considered:

   - A two-rectangle feature: the difference between the sum of the pixel values of two adjacent rectangular windows;
   - A three-rectangle feature: the difference between sum of the pixel values in the extreme rectangles and the sum of the pixel values in the middle rectangle;and
   - A four-rectangle feature: the difference between sum of the pixel values in the rectangles that constitute the main and off diagonals in a 2×2 set of rectangles.

2. AdaBoost based learning algorithm [20]. In a standard 24×24 sub-window, there are 45396 possible Haar-like features. The task of the AdaBoost algorithm is to pick a few hundred features and assign weights to each using a set of training images. Face detection problem is now reduced to computing the weighted sum of the chosen rectangle features and applying a threshold.

3. Cascade architecture. In order to achieve increased detection performance as well as radically reduce computation time, the strong classifiers are arranged in a cascade in order of complexity.

In our implementation, after detecting the face region in the image, detectors of the right and left eyes and of the mouth are applied to it in a sequence. For better accuracy in eye and mouth detection, the eye detectors are applied to the

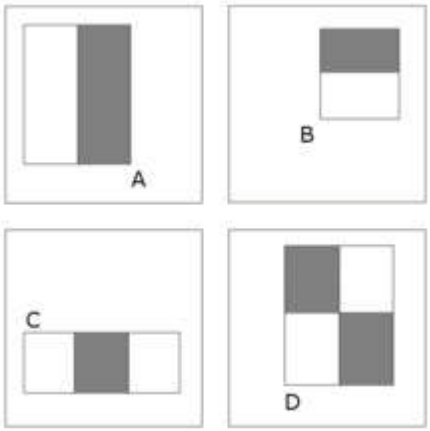


Figure 6.2: The Haar-like features used in the Viola-Jones' method [81].

upper part of the face region, while the mouth detector is applied to its lower part. When all the three facial components are detected, the system computes the coordinates of centers of the detected eyes and mouth, based on the detected facial components. These center points are marked in the uploaded images for the user. Figure 6.3 shows some typical results of such an operation.

The user either decides to accept the marked points or provide, manually, more accurate centers of the eyes and of the mouth, which are critical for face normalization (see below) and hence for an improved performance of expression recognition in the final stage.

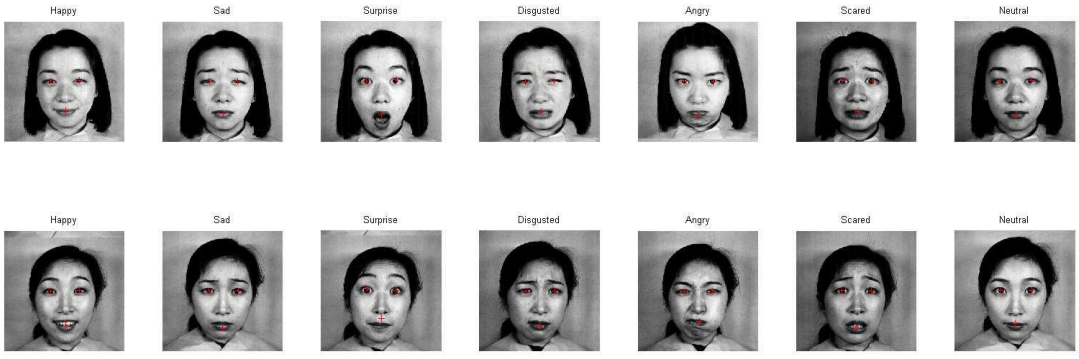


Figure 6.3: The results of using eyes and mouth detection on sample images from the JAFFE database.

6.3 Face Normalization

This is performed to achieve uniformity in size, position and illumination condition. Note that we consider facial expression recognition from only still frontal images, which are possibly subject to moderate variations in pose and illumination. We adopt affine transformation and retinex-based illumination normalization to achieve fast and efficient face normalization, as explained below.

6.3.1 Affine Transformation for Pose Normalization

The eyes and mouth of the uploaded image are to lie at the same positions as those in the training images. Mathematically, for each pixel (x_i, y_i) in the uploaded image, an affine transformation is to be determined such that all the corresponding projected pixels (px_i, py_i) are in the desired positions, as a solution to the following equation:

$$\begin{aligned} px_i &= a_1 * x_i + a_2 * y_i + a_3 \\ py_i &= a_4 * x_i + a_5 * y_i + a_6, \end{aligned} \tag{6.1}$$

where a_i for $i = 1, 2, \cdots, 6$ are the unknown parameters. Its solution requires at least three corresponding pairs of points in both the uploaded and training images. In our implementation, the centers of eyes and mouth are used to obtain the parameters a_i. Figure 6.4(a) shows the centers of eyes and mouth of the mean face of the training images, while Figure 6.4(b) shows those of a sample input image, having a moderate pose variation. All the points in the uploaded image are mapped using the affine transformation to obtain a new image, in which the positions of the centers of the eyes and mouth match those of the average of the training images. Figure 6.4(c) shows the affine-transformed version of Figure 6.4(b).

The face region in the affine transformed image is next cropped to a 184×152 image, to match the size of training images. At the end of the above operations on the uploaded image, we arrive at a pose- and size-normalized image.

(a) Centers of eyes and mouth of the mean face of the training images

(b) Centers of eyes and mouth of one sample input image.

(c) Affine transformed image.

Figure 6.4: Example of pose normalization.

6.3.2 Retinex Based Illumination Normalization

Since the uploaded image may contain illumination variations, we normalize it by adopting the Retinex [45], an image enhancement algorithm that provides a high level of dynamic range compression.

The basic Retinex, referred to as the single scale retinex (SSR) [37], is defined for any point (x, y) in an image by the following equation:

$$R_i(x,y) = logI_i(x,y) - log[F(x,y) \otimes I_i(x,y)], \tag{6.2}$$

where $I_i(x,y)$ is the image intensity in the i^{th} channel; $R_i(x,y)$ is the retinex output; $\otimes$ represents the convolution operator; and $F(x,y) = Kexp(-\frac{x^2+y^2}{c^2})$ is the Gaussian kernel function with effective width c, the scale of the Retinex; and K is the scale factor, obtained by setting $\iint F(x,y)dxdy = 1$. Note that a small value of c leads to good dynamic range compression, while its large value leads to good tonal rendition for a color image. Figure 6.5 shows sample SSR images with different scales.

To obtain a better balance of dynamic compression and color rendition, multi-scale retinex (MSR) [36] has been proposed by combining several SSRs:

$$R_{MSR_i} = \sum_{n=1}^{N} w_n\{logI_i(x,y) - log[F_n(x,y) \otimes I_i(x,y)]\}, \tag{6.3}$$

Figure 6.5: SSR images with different scales.

where N is the number of scales; and w_n is the weight of the n^{th} scale. In our implementation, $N = 3$, $w_n = \frac{1}{3}$ and $c = 10, 40, 120$. Figure 6.6 shows sample MSR images using the above parameters.

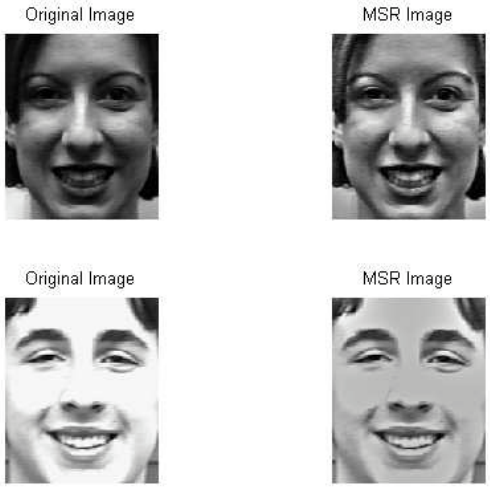


Figure 6.6: MSR images with empirical parameters.

Now the face region of the uploaded image is ready for facial expression recognition, which is the key procedure of the system.

6.4 Local Gabor Feature Based Facial Expression Recognition

The facial expression recognition component of this system is based on the radial encoded local Gabor features and classifier synthesis algorithm, as described in Chapter 5. Recall that the general procedure is as follows:

- An input image is divided into several local blocks and a set of Gabor filters with different orientations and scales are applied to every local block;
- Radial encoding strategy is utilized to downsample the Gabor outputs and for every local block the encoded Gabor outputs from the different orientations but the same scale are grouped together as the local features;
- A set of local classifiers are assigned to generate local decisions based on the local features and the local decisions are concatenated to form the intermediate features.
- A global classifier is assigned to make the final decision according to the intermediate features.

In this chapter, we focus on addressing the implementation issues with a goal of balancing the recognition performance and operating speed.

6.4.1 The Training Database

We use the combination of the JAFFE and CK databases as the training set. As discussed in Chapter 5, we select 2581 images with clear facial expressions from the CK database and similarly 213 images from the JAFFE database. These images are form 104 different persons, including American, Latino, European and Asian. All the training images are normalized using pose and illumination normalization methods and the size is fixed to be 184×152.

6.4.2 The Number of Local Blocks

Although simulation results in Chapter 5 show that 49 local blocks are the best trade-off between performance and computational load, we discover that it is still inefficient for a real-time application in terms of respond time to user. To reduce the computational time without significant degradation in performance, 25 local blocks are now considered in this chapter when implementing the system.

6.4.3 Support Vector Machine (SVM)

The k-nearest neighbor (KNN) classifier used in Chapter 5 is the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors. Especially when $k = 1$, the object is simply assigned to the class of its nearest neighbor in the feature space. Thus the recognition performance using KNN may not be the optimal one compared to other advanced classifiers. In addition, the system has to store all the related information of the training images for computing the distances between a new input image and the training images. These distances are used to locate the nearest neighbor of the input image such that its expression can be classified. Note that in our implementation there are 2794 training images, so the storage of the training features is memory-consuming. Therefore, KNN cannot fulfill the requirement of the real-time application. Instead, we use the support vector machine (SVM) [10] as the classifier. The original SVM is proposed to find the optimal hyperplane which separates two different classes. Since 7 basic facial expressions are involved in this thesis, we adopt the multi-class SVM with one-against-all strategy, in which the classifier with the highest output assigns the class to the test data. Notice that the local KNN classifier in the Chapter 5 is modified to produce estimated probabilities of each class and these probabilities are further used to form the intermediate features. Obviously, the SVM needs to be modified to produce similar outputs:

- For a given test input vector, compute SVM scores for each class $[S_1, S_2, ..., S_C]$, where C is the number of classes;

- Estimate the probabilities for each class using the following indication function:
$$P_i = \frac{exp(S_i)}{\sum_{k=1}^{C} exp(S_k)},\ i = 1, 2, \cdots, C, \tag{6.4}$$
and form the output vector using $P_1,\ P_2,\ \cdots,\ P_C$.

Here, the exponential function is chosen as the indication function for estimating the probability for each class since larger SVM score of certain class indicates the test sample more likely belongs to that class.

6.4.4 Other Related Parameters

The remaining parameters to be decided for the application are listed as follows:

1. Parameters of Gabor filters;
2. Resolution of radial grid for radial encoding;
3. Number of extracted components N_c and the regularization factor β.

We use the same configurations for Gabor filters as discussed in Chapter 5: (i) eight orientations (from 0 to $7/8\pi$ in uniform steps of $1/8\pi$) in order to capture subtle changes of facial components, and (ii) three filter sizes ($\mathcal{S}$ = 11, 20 and 29) in order to determine the σ and λ according to the following formulas [64]:

$$\sigma = 0.0036\ \mathcal{S}^2 + 0.35\ \mathcal{S} + 0.18, \quad \text{and} \quad \lambda = \frac{\sigma}{0.8}. \tag{6.5}$$

The remaining two parameters, φ, and γ, are set to 0 and 0.5, respectively.

Similarly, for radial encoding, we choose a grid of size 18×5 (angular *vs* radial) such that the innermost sector contains at least one pixel, thereby resulting in a feature matrix of size, 18×5 for each Gabor filtered local block.

Moreover, we use the same cross-validation strategy to determine β and N_c:

- Divide the training set (CK and JAFFE databases) into 10 groups according to the identity information.
- Use one group as the validation set, and the other groups to train the system. Repeat until every group has been chosen as a validation set once. Average the recognition accuracy of validation sets.

Table 6.1: Recognition accuracies (%) of the system on the generalization test with different configurations.

	TFEID	YALE	FEED
KNN, 49 blocks, without Face Normalization	61.94	60.66	61.43
KNN, 49 blocks, with Face Normalization	75.75	66.67	69.66
SVM, 25 blocks, with Face Normalization	82.84	63.33	71.30

- Use the SPSA algorithm to find the optimal values of β and N_c for validation sets.

By using the validation set, β and N_c are determined to be 1.1 and 18, respectively.

Table 6.1 shows the simulation results of the proposed system on TFEID, YALE and FEED databases. For comparison, the first line contains the results of using KNN without face normalization while the second line contains the results of using KNN with face normalization. The significant improvement in recognition accuracy indicates the advantage of face normalization. Moreover, the third line contains the results of using SVM with face normalization and 25 local blocks. Also, the improvement in recognition accuracy indicates SVM classifier outperforms KNN.

6.5 Experimental Test of the Facial Expression System

Based on the proposed framework, we develop a web-based application.[1] Figure 6.7 shows a snapshot of the user interface (UI) of the proposed system. When the test image is uploaded, the system checks whether it contains a human face. As shown in Figure 6.8 and 6.9, if there is no human face, the UI asks the user to upload a face image. If the test image contains a human face, the system then

[1]Check "http://137.132.165.17/default.aspx".

detects locations of the centers of the eyes and mouth of the input image, and displays the results to the user, as shown in Figure 6.10. The user decides whether the detected centers of eyes and mouth are accurate enough for recognizing expression. The user can also use UI to mark, manually, the centers of eyes and mouth of the test image if the system fails to detect eyes and mouth accurately, as shown in Figure 6.11 and Figure 6.12. Once the user presses the "Confirm" button in the UI, the system starts to process the test image, and after about 10 seconds it displays the final recognition results via UI, as shown in Figure 6.13.

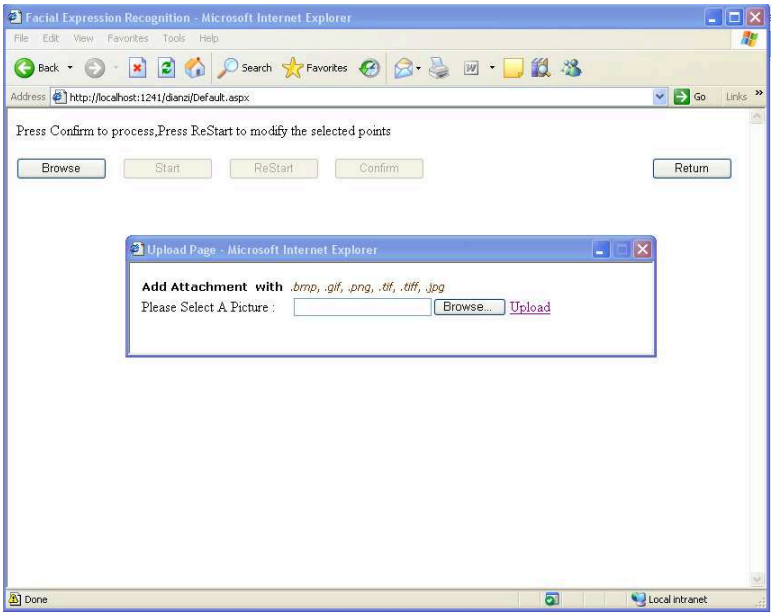


Figure 6.7: The snapshot of the UI of the proposed system.

To test the performance of the proposed system, we collect totally 70 facial images with 7 basic expressions (10 images per class) from the internet. Figure 6.14 shows all the test image, and from the top to bottom, each line contains images with happy, sad, surprise, disgusted, angry, scared and neutral expression, respectively. Table 6.2 shows the recognition performance of the proposed system. From the recognition results, we observe that the average accuracy is 72.86%, which is close to the result of generalization test discussed in the previous section. Moreover, all the images with happy and surprise expressions are

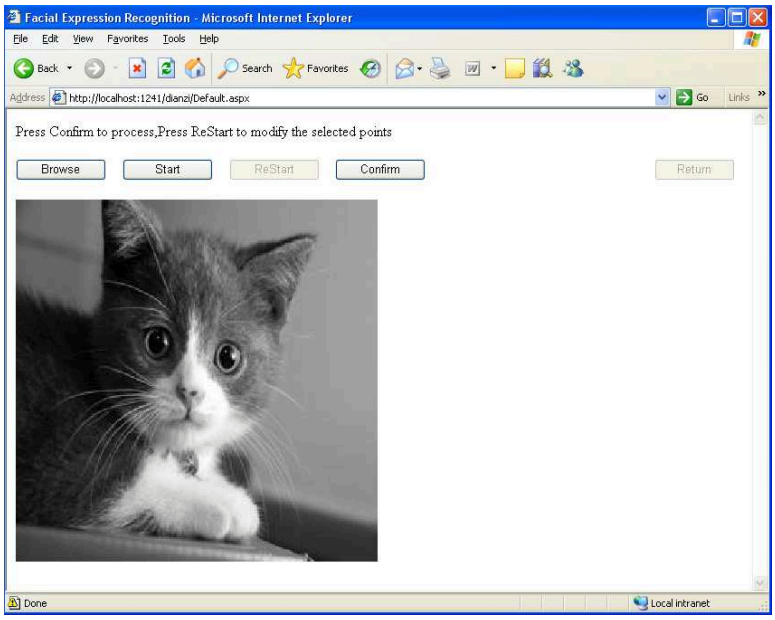


Figure 6.8: The uploaded image contains a cat face rather than a human face.

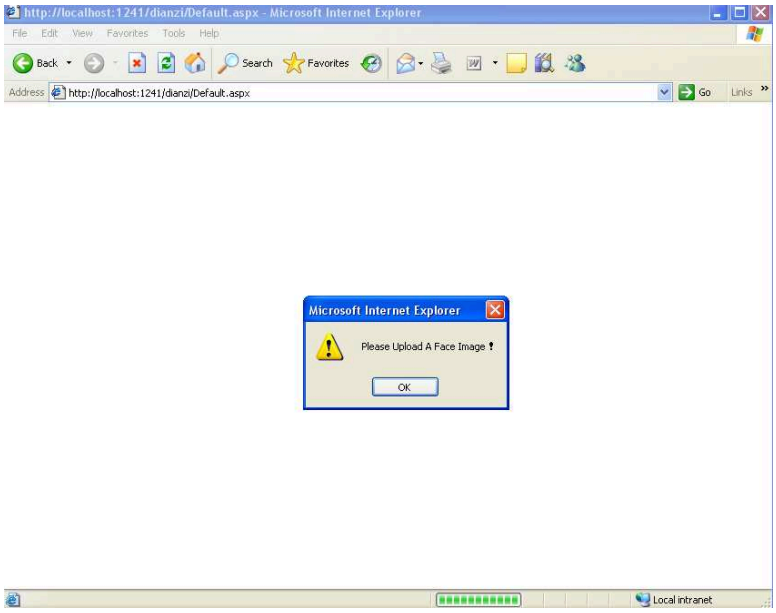


Figure 6.9: The UI asks the user to upload a human face.

correctly recognized while images with scared expression are difficult to recognize. All the misclassified images with scared expression are recognize as surprise

Figure 6.10: The detected eyes and mouth of a test image.

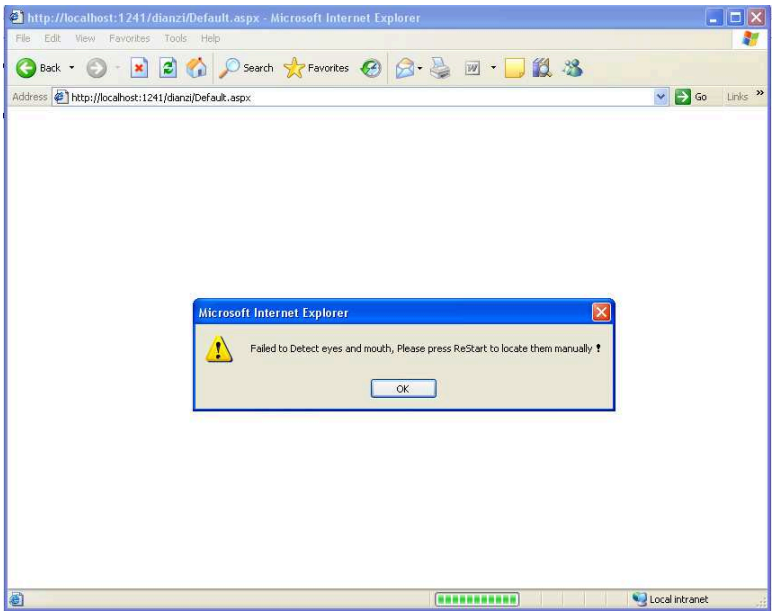


Figure 6.11: The UI shows that the system fails to detect eyes and mouth of a test image.

expression, for example, as shown in Figure 6.15. This is due to the fact that

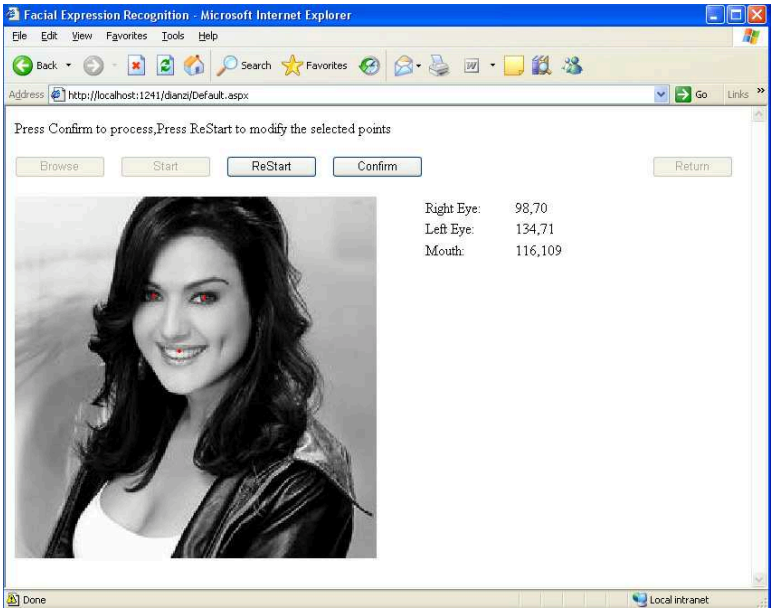


Figure 6.12: The user uses the UI to specify the centers of eyes and mouth of a test image.

Figure 6.13: The UI shows the final recognition result of a test image.

most of scared images we collect from internet actually contain mixed expressions

of surprise and scared. However, for happy expression, the proposed system can successfully recognize the images even with occlusions, as shown in Figure 6.16 and 6.17. Notice that all the test images used here are randomly collected from the internet and they are quite different from the training images in terms of illumination condition, pose and especially the identity of the person. Taking all these affects into account, it demonstrates the outstanding performance of the proposed system on facial expression recognition from still images from novel persons. Figure 6.18 - 6.24 show more results of the proposed system on the test images from the internet.

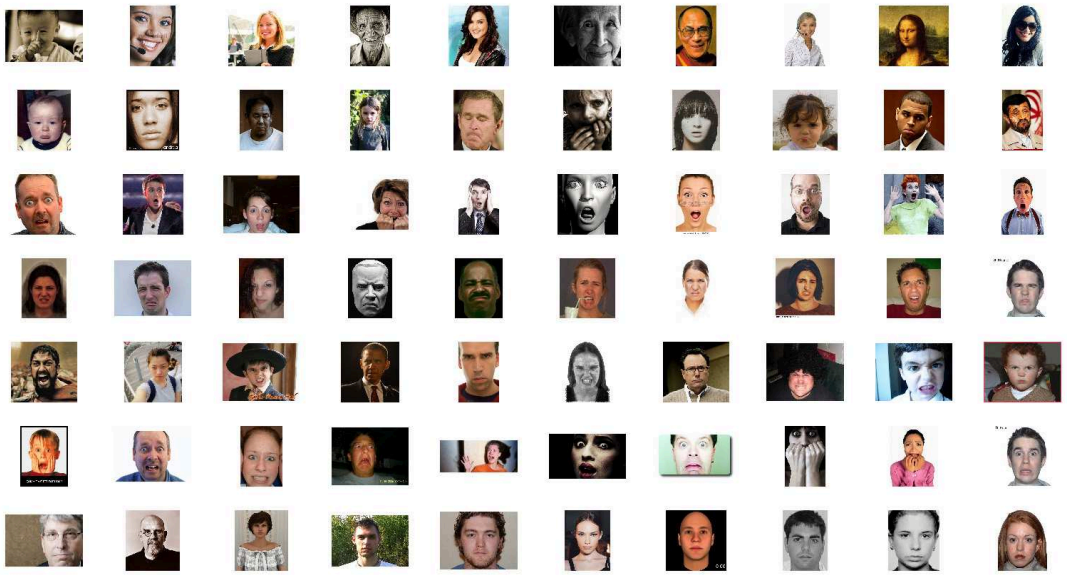


Figure 6.14: The test images collected from the internet.

Table 6.2: Recognition results (%) of the proposed system on the test images from internet.

Happy	Sad	Surprise	Disgusted	Angry	Scared	Neutral	Average
100	60	100	70	60	50	70	72.86

Figure 6.15: The scared expression is misclassified as surprise.

Figure 6.16: The happy image with mouth occlusion.

Figure 6.17: The happy image with eye occlusion.

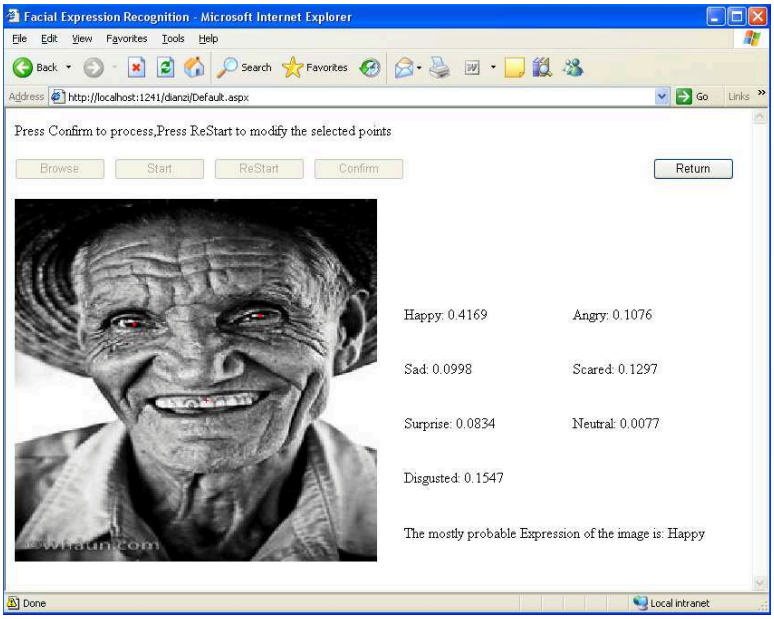


Figure 6.18: The recognized happy image from the internet.

6.6 Summary

In this chapter, we have successfully implemented an efficient web-based facial expression recognition system based on our proposed hierarchical facial expres-

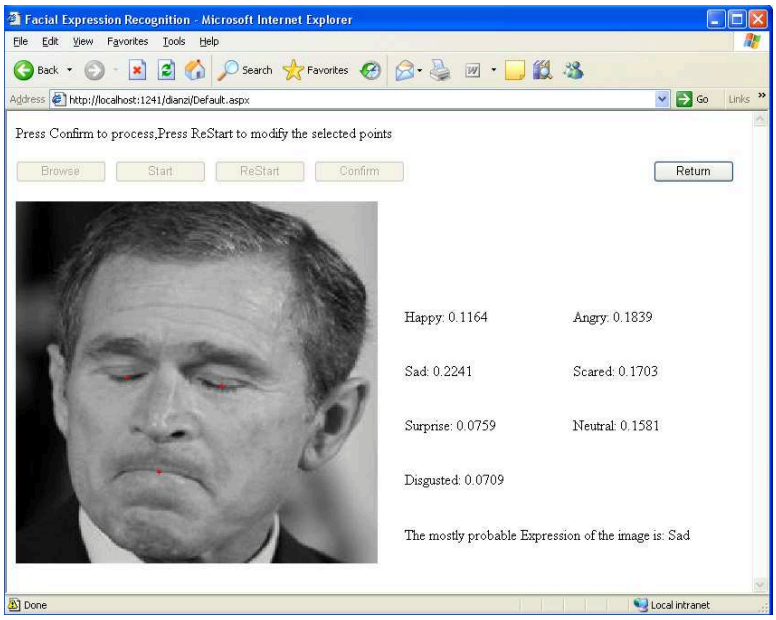


Figure 6.19: The recognized sad image from the internet.

Figure 6.20: The recognized surprise image from the internet.

sion recognition framework, employing face and facial components detection algorithms as well as a face normalization technique. Experimental results show that

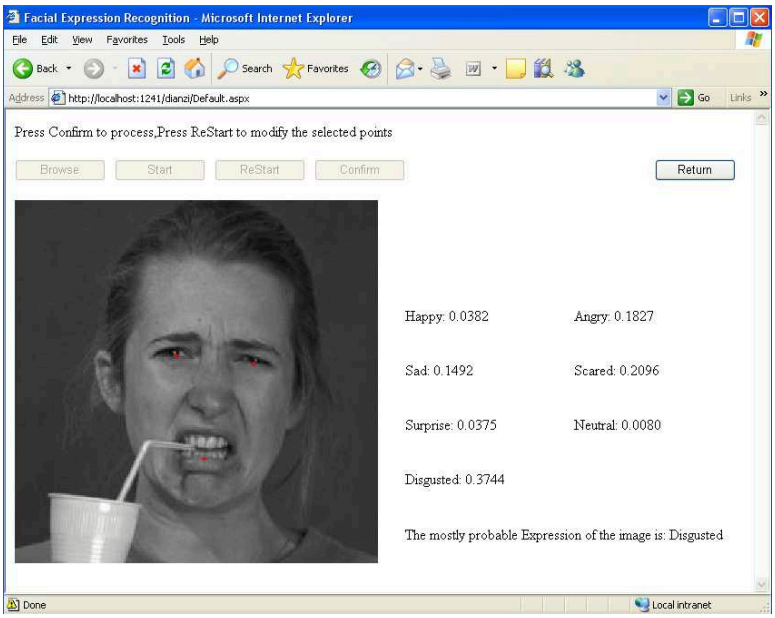


Figure 6.21: The recognized disgusted image from the internet.

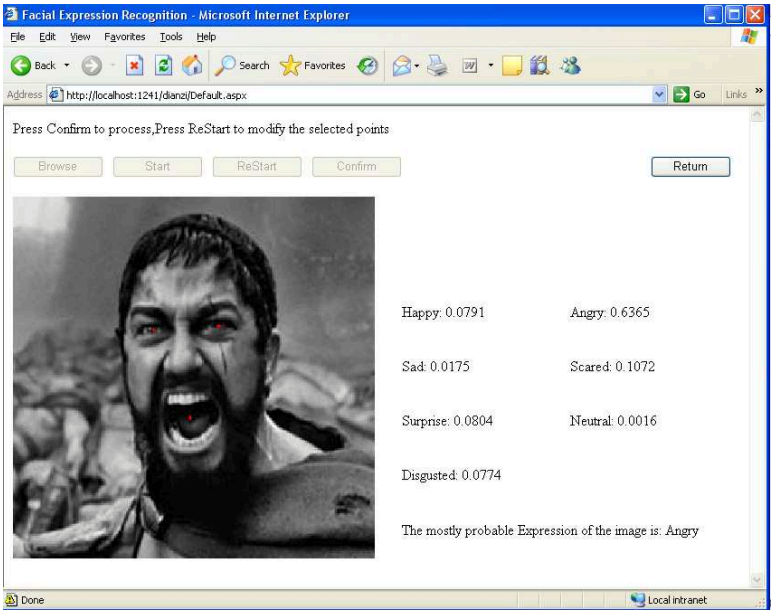


Figure 6.22: The recognized angry image from the internet.

the proposed system can automatically recognize facial expressions of a facial image from the internet with high recognition accuracy, thereby pointing the way to

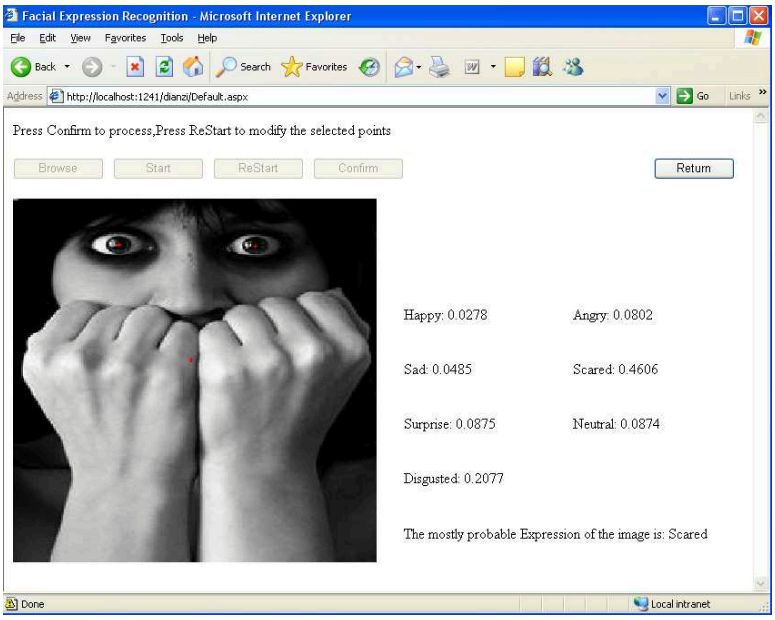


Figure 6.23: The recognized scared image from the internet.

Figure 6.24: The recognized neutral image from the internet.

the development of a robust and stable facial expression recognition system with improved recognition accuracy when recognizing expressions of novel persons.

Chapter 7

Conclusions

Humans can effortlessly recognize facial expressions, which mirror emotions, and respond to them appropriately. Since this cognitive ability, which is one aspect of human intelligence, is not completely understood, attempts are being made to design machines to recognize facial expressions in the hope that the implemented algorithm provides an insight into human intelligence. It has been found that such machines can barely recognize facial expressions of the class of humans whose images have been used for training such machines but not of those not belonging to that class (i.e., the class of strangers or "novel" persons). The implication is that facial expression is normally correlated with identity, and variations in identity affect the (machine) recognition of expressions. Therefore, there is a need to develop a "person-independent" expression recognition system, i.e., a system which is also applicable to novel faces. To this end, the thesis proposes a new framework which combines the characteristics of the human visual system with statistical pattern recognition techniques.

7.1 Main Contributions

Motivated by the contour-extraction characteristics of retinal ganglion cells, we have proposed, in Chapter 2, an efficient algorithm for recognizing facial expressions, using the contours of face (and its parts) as features. For both person-dependent and person-independent recognition of expressions, it is found that

these features lead to the algorithm's good performance which compares favorably with the accuracy of recognition of expressions, in the same images, by humans. An important feature of this algorithm is that it suggests, facial contours and its components, extracted by using the level-set method, have been here, for the first time, successfully employed in facial expression recognition, thereby demonstrating that they (i.e., facial contours) contain information about facial expressions, and are, therefore, biologically plausible features in the human perception of facial expressions.

Based on recent physiological findings in the human brain related to face processing, the power of a biologically inspired approach for significantly improving expression recognition accuracy is explored in Chapter 3 by combining the HMAX model with local methods and face processing units. The improvement may be attributed to the elegant structure of local methods which model face-selective cells in the FFA of the human visual system (HVS). Experimental results show that the local classifier combination method, using PCA along with FLD analysis, performs better than classical classifier combination rules, such as Borda count and decision template. The underlying strategy is the design of a new framework for expression recognition by exploiting the hierarchal structure of the HVS.

In an attempt to simulate the expression-selective cells in the STS of the HVS, a composite orthonormal basis (COB) algorithm is, proposed in Chapter 4. It is found that the COB can extract, from the face images, an expression subspace with the identity information removed as much as possible. This sub-space corresponds to the global features of a facial expression. When combined with local methods, the COB decouples expression from identity, and results in outstanding expression recognition performance when applied to different databases. This demonstrates the power of fusing a statistical COB-based approach with (bio-inspired) local methods on person-independent facial expression recognition.

By way of further exploring bio-inspired models and statistical techniques, radial encoded Gabor features and a local classifier synthesis are combined to form a new hybrid framework for expression recognition in Chapter 5. The retinotopic mapping structure of the HVS is modeled by the radial encoding of Gabor features, thereby effectively downsampling the outputs of local Gabor filters as applied to local patches of input images. Local classifiers are then employed to make

the local decisions, which are integrated to form intermediate features for representing facial expressions globally. Experimental results show that the encoded features are discriminatory enough to outperform classical statistical techniques that invoke Gabor jets, based on fiducial points and a uniform downsampling method. Recognition accuracies with respect to standard individual databases are significantly better than those in the literature. Furthermore, the proposed system can also recognize expressions in most of the images from an altogether different database, which seems to be the first satisfactory cross-database recognition performance. With the help of appropriate tests, the proposed framework has also been shown to be robust to corrupted data and to missing information.

Finally, in Chapter 6, a real-time web-based application of facial expression recognition system, based on the hybrid framework (of Chapter 5), is implemented, in which, in order to be useful for practical applications, algorithms for (i) detecting face and its components; and (ii) face normalization are integrated. For classification, the SVM is employed to facilitate real-time processing. Experimental results demonstrate that the proposed system can automatically, and also highly accurately, recognize the expression of an image, uploaded from the internet. This system is expected to shed light on developing a robust and stable system to recognize expressions of novel persons more accurately.

7.2 Future Research Directions

In this section, we list several future research directions that are related to our work.

1. In the local feature integration stage of our proposed scheme, we use a combination of classifiers. This is different from the human vision system, which produces intermediate features by combining low-level features. However, since the mechanism of feature combination in the human brain is still not known clearly, we resort to a statistical approach. The proposed classifier combination is one possible solution for integrating local features. In order to produce more discriminating intermediate features for improving

final recognition performance, there is a need for a new strategy to combine features.

2. In the proposed framework of facial expression recognition, we use a supervised learning strategy in both the low- and high-level layers. This seems to be inconsistent with what is known about the HVS. Physiological researches indicate that the HVS involves unsupervised learning in the low-level layers, such as V1 and V2. Such a learning mechanism helps cells in V1 and V2, which have a similar functional structures, to integrate low-level features into intermediate-level features. On the other hand, in the high-level layers of the HVS, supervised learning plays an important role in extracting discriminating features to recognize objects from intermediate-level features. An interesting problem is whether a combination of unsupervised and supervised learning in such a hierarchical manner contributes to improving the performance of expression recognition.

3. It has been found that the web-based facial expression recognition system is sensitive to the coordinates of centers of eyes and mouth. If an uploaded image contains a face region with low resolution, a minor shift in centers of eyes and mouth leads to a significant decrease in the accuracy of expression recognition. Since all the algorithms have been implemented in MATLAB, expression recognition is slow for a real-time application. It is desirable optimize the web-based expression recognition system for real-time applications.

4. The present study has considered expression recognition only from static images. For video sequences, a new approach is needed since the motion of specific facial regions seems to provide additional features characterizing various expressions which can be exploited. Spontaneous expressions can also be treated as dynamic for which motion features are crucial. It is likely that Gabor features will not play any tangible role in their recognition, and further research is needed to extract new features. An additional challenge is how to identify the optical flow corresponding to the dynamics of an

expression. In addition, what is an proper feature encoding strategy to reduce the computational load to achieve real-time (expression) recognition?

To conclude, automatic person-independent facial expression recognition is still *largely* an unresolved and challenging problem. A new, bio-inspired machine paradigm, which incorporates the essential features of the HVS in a statistical framework, is needed to enhance the recognition capability of present-day machines to a level comparable to that of human beings. The thesis represents a step in that direction.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face Recognition With Local Binary Patterns. *ECCV*, pages 469–481, 2004. 12

[2] P. Aleksic and A. Katsaggelos. Automatic Facial Expression Recognition Using Facial Animation Parameters and Multi-stream HMMS. *IEEE Trans. Information Forensics and Security*, 1:3–11, 2006. 81

[3] M. Bartlett, G. Littlewort, I. Fasel, and R. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Application to Human Computer Interaction. In *Proc. CVPR Workshop Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003. 59, 80, 81

[4] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, July 1997. 5, 66

[5] B.Fasel and J.Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36:259–275, 2003. 43

[6] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with The OpenCV Library*. O'Relly, 2008. 84

[7] V. Bruce and A. Young. Understanding Face Recognition. *The British Journal of Psychology*, 77(3):305–327, 1986. 1

[8] L. Chen and Y. Yen. Taiwanese Facial Expression Image Database, 2007. URL `http://bml.ym.edu.tw/~download/html`. 35

[9] M. Connolly and D. Van Essen. The Represetation of The Visual Field in Parvicellular and Magnocellular Layers of The Lateral Geniculate Nucleus in The Macaque Monkey. *Journal of Comparative Neurology*, 226(4):544–564, 1984. 63

[10] C. Cortes and V. Vapnik. Support-Vector Metworks. *Machine Learning*, 20(3): 273–297, 1995. 90

[11] N. Costen, T. Cootes, G. Edwards, and C. Taylor. Automatic Extraction of The Face Identity-subspace. *Image Vision Computing*, 20:319–329, 2002. 1

[12] J. Daugman. Uncertainty Relations for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-dimensional Visual Cortical Filters,. *Journal of the Optical Society of America*, 2:1160–1169, 1985. 9

[13] H. Deng, L. Jin, L. Zhen, and J. Huang. A New Facial Expression Recognition Method Based on Local Gabor Filter Bank and PCA plus LDA. *International Journal of Information Technology*, 11(11):86–96, 2005. 11

[14] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, New York, 2nd edition, 2001. ISBN 0471056693. 2, 5, 42

[15] P. Ekman. An Argument for Basic Emotions. *Cognition and Emotion*, 6:169–200, 1992. 21

[16] K. Etemad and R. Chellappa. Discriminant Analysis for Recognition of Human Face Images. *J. Opt. Soc. Am. A*, 14(8):1724–1733, Aug 1997. 5

[17] T. Ezzat and T. Poggio. Facial Analysis and Synthesis Using Image-based Models. In *Proc. Second International Conference on Automatic Face and Gesture Recognition*, pages 116–121, Vermont, USA, 1996. 22

[18] X. Feng, A. Hadid, and M. Pietikäinen. A Coarse-to-fine Classification Sheme for Facial Expression Recognition. In *Proc. First International Conference on Image Analysis and Recognition*, pages 668–675, Porto, Portugal, 2004. 59, 80

[19] R. Fisher. The Use of Multiple Measures in Taxonomic Problems. *Ann. Eugenics*, 7:179–188, 1936. 4

[20] Y. Freund and R. Schapire. A Decision-theroretic Generalization of On-line Learning and An Application to Boosting. In *Proc. The Second European Conference on Computational Learning Theory*, volume 904, pages 23–37, 1995. 84

[21] K. Fukunaga. *Statistical Pattern Recognition*. Adcademic Press, 1990. 5

[22] K. Fukushima. Neocognitron: A Self-organizing Neural Network Model for A Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 36(4):93–202, 1980. 8

[23] M. Ganesh and Y. Venkatesh. Efficient Classification by Neural Networks Using Encoded Patterns. *Electronics Letters*, 31:400–403, 1994. 25

[24] M. Ganesh and Y. Venkatesh. Modified Neocognitron for Improved 2-D Pattern Recognition. In *IEEE Proceedings-Vis.Image and Signal Processing*, volume 143, pages 31–40, 1996. 8

[25] X. Geng and Y. Zhang. Facial Expression Recognition Based on The Difference of Statistical Features. In *Proc. International Conference on Singal Processing*, pages 16–20, 2006. 80

[26] M. Goodale and A. Milner. Separate Pathways for Perception and Action. *Trends in Neuroscience*, 15(1):20–25, 1992. 6

[27] G. Gottumukkal and V. Asari. An Improved Face Recognition Technique Based on Modular PCA Approach. *Pattern Recognition Letters*, 25(4):429–436, 2004. 11

[28] W. Gu, Y. Venkatesh, and C. Xiang. A Novel Application of Self-organizing Network for Facial Expression Recognition From Radial Encoded Contours. *Soft Computing*, Online First™, 2009. 64

[29] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, USA, 2001. 66

[30] L. He, C. Zou, L. Zhao, and D. Hu. An Enhanced LBP Feature Based on Facial Expression Recognition. In *Proc. IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 3300–3303, Shanghai, China, 2005. 59, 80

[31] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face Recognition: Component-based Versus Global Approaches. *Comput. Vis. Image Understand.*, 91(1):6–12, 2003. 12

[32] D. Huang, C. Xiang, and S. Ge. Feature Extraction for Face Recognition Using Recursive Bayesian Linear Discriminant. In *Proc. International Symposium on Image and Signal Processing and Analysis*, pages 356–361, Istanbul, Turkey, 2007. 37

[33] D. Hubel. *Eye, Brain and Vision (Scientific American Library, Vol.22)*. W.H.Freeman, New York, USA, 1988. 6, 62

[34] D. Hubel and T. Wiesel. *Brain and Visual Perception, The Story of a 25-Year Collaboration*. Oxford, New York, USA, 2005. 7

[35] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, New Jersey, USA, 1998. 65

[36] D. Jobson, Z. Rahman, and G. Woodell. A Multi-scale Retinex for Bridging The Gap between Color Images and The Human Observation of Scenes. *IEEE Trans. Image Processing*, 1997. 87

[37] D. Jobson, Z. Rahman, and G. Woodell. Retinex Processing for Automatic Image Enhancement. In *Proc. SPIE Symposium on Electronic Imaging*, 2002. 87

[38] J. Jones and L. Palmer. An Evaluation of The Two-dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. *Journal of Neurophysiology*, 6:1233–1258, 1987. 9

[39] M. Kamachi, M. Lyons, and J. Gyoba. The Japanese Female Facial Expression (JAFFE) Database. URL `http://www.kasrl.org/jaffe.html`. 21

[40] T. Kanade, J. Cohn, and Y. Tian. Comprehensive Database For Facial Expression Analysis. In *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, pages 46–53, 2000. 54

[41] M. Kirby and L. Sirovich. Application of The Karhumen-Loève procedure for The Characterization of Human Faces. *IEEE Trans. Pattern. Anal. Mach. Intell.*, 12: 103–108, 1990. 3

[42] T. Kohonen. *Self-Organizing Map*. Springer-Verlag, Berlin, Germany, 2nd edition, 1995. 21, 27, 55

[43] I. Kotsia, S. Zafeiriou, and I. Pitas. Novel Class of Multiclass Classifiers Based on The Minimization of Within-class-variance. *IEEE Trans. Neural Networks*, 20(1): 14–34, 2009. 73, 74, 76

[44] M. Kyperountas, A. Tefas, and I. Pitas. Salient Feature and Reliable Classifier Selection for Facial Expression Classification. *Pattern Recognition*, 43(4):972–986, 2010. 80

[45] E. Land. An Alternative Technique for The Computation of The Designator in The Retinex Theory of Color Vision. In *Proc. Natl Acad Sci*, volume 83, pages 3078–3080, 1986. 87

[46] C. Li, C. Xu, C. Gui, and M. Fox. Level Set Evolution without Reinitialization: A New Variational Formulation. In *Proc. IEEE Computer Society International Conference on Computer Vision and Pattern Recognition*, pages 430–436, San Diego, USA, 2005. 24

[47] Z. Li, J. Imai, and M. Kaneko. Facial Expression Recognition Using Facial-component-based Bag of Words and PHOG Descriptors. *Information and Media Technologies*, 5(3):1003–1009, 2010. 80, 81

[48] L.I.Kuncheva. *Combining Pattern Classifiers, Methods and Algorithms*. Wiley Interscience, New York, USA, 2005. 44

[49] J. L.I.Kuncheva and R.Duin. Decision Templates for Multiple Classifier Fusion: An Experimental Comparison. *Pattern Recognition*, 34(2):299–314, 2001. 44

[50] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of Facical Expression Extracted Automatically from Video. In *Proc. IEEE Workshop Face Processing in Video*, 2004. 81

[51] J. Louie. A Biological Model of Object Recognition with Feature Learning. Technical report, MIT, Massachusetts, USA, 2003. 14, 15

[52] M. Lyons, J. Budynek, and S. Akamatsu. Automatic Classification of Single Facial Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21:1357–1362, 1999. 5, 37, 80

[53] G. J. Mclachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, 1992. 5

[54] M. Moller. A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. *Neural Networks*, 6:525–533, 525-533. 34

[55] J. Nolte. *The Human Brain: An Introduction to Its Functional Anatomy*. Mosby, St.Louis, 5th edition, 2002. 7

[56] S. Osher and J. Sethian. Fronts Propagating With Curvature-dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations. *Journal of Computation Physics*, 79:12–49, 1988. 24, 25

[57] P. Padgett and G. Cottrell. Representing Face Image for Emotion Classification. *Advances in Neural Information Processing Systems*, 9:894–900, 1996. 3

[58] K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(6):559C572, 1901. 3

[59] A. Pentland, B. Moghaddam, and T. Starner. View-based and Modular Eigenspaces for Face Recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 84–91, 1994. 3, 11

[60] M. Poetzsch, N. Krueger, and C. von der Malsburg. Improving Object Recognition by Transforming Gabor Filter Responses. *Network:Computation in Neural Systems*, 7:341–347, 1996. 10

[61] M. Riesenhuber and T. Poggio. Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999. vii, 8, 12, 14, 25

[62] G. Rogova. Combining The Results of Several Neural Network Classifiers. *Neural Networks*, 7(5):777–781, 1994. 67

[63] A. Samal and P. Iyengar. Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey. *Pattern Recognition*, 25(1):65–77, 1992. 1

[64] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A Theory of Object Recognition: Computations and Circuits in The Feedforward Path of The Ventral Stream in Primate Visual Cortx, Technical Report. Technical report, MIT, Massachusetts, USA, 2005. vii, 14, 15, 41, 62, 63, 91

[65] C. Shan, S. Gong, and P. McOwan. Robust Facial Expression Recognition Using Local Binary Patterns. In *Proc. IEEE Int'l Conf. Image Processing*, pages 370–373, 2005. 59, 80, 81

[66] Y. Shinohara and N. Otsu. Facial Expression Recognition Using Fisher Weight Maps. In *Proc. Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 499–504, 2004. 5, 20, 37, 80

[67] M. Simon. *Facial Expression: A Visual Reference for Artists*. Watson-Guptill, New York, USA, 2005. vii, 20, 21

[68] R. Snowden, P. Thompson, and T. Troscianko. *Basic Vision: An introduction to visual perception*. Oxford, New York, USA, 2006. 20

[69] J. Spall. Implementation of The Simultaneous Perturbation Algorithm for Stochastic Optimization. *IEEE Trans. Aerospace and Electronic Systems*, 34:817–823, 1998. 57

[70] B. Sumengen. A Matlab Toolbox Implementing Level Set Methods, 2004. URL http://barissumengen.com/level_set_methods. 24

[71] J. Sun, Q. Zhuo, and W. Wang. An Improved Facial Expression Recognition Method. In *Proc. Advances in Multimodal Interfaces, ICMI 2000*, pages 215–221, 2000. 5

[72] D. Swets and J. Weng. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836, Aug 1996. 5

[73] K. Tan and S. Chen. Adaptively Weighted Sub-pattern PCA for Face Recognition. *Neurocomputing*, 64:505–511, 2005. 12

[74] Y. Tian. Evaluation of Face Resolution for Expression Analysis. In *IEEE Workshop Face Processing in Video*, 2004. 81

[75] R. Tootell, M. Silerman, E. Switkes, and R. De Valois. Deoxyglucose Analysis of Retinotopic Organization in Primates. *Science*, 218:902–904, 1984. 63

[76] D. Tsao. A Dedicated System for Processing Faces. *Science*, 314:72–73, 2006. 8, 11, 43, 52, 65

[77] D. Tsao, W. Freiwald, R. Tootell, and M. Livingstone. A Cortical Region Consisting Entirely of Face-selective Cells. *Science*, 311:670–674, 2006. 8, 11, 43, 52, 65

[78] D. Tsao, N. Schweers, S. Moeller, and W. Freiwald. Patches of Face-selective Cortex in The Macaque Frontal Lobe. *Nature Neuroscience*, 11(8):877–879, 2008. 8, 11, 43, 52, 65

[79] M. Turk and A. Pentland. Eigenfaces for Recognition. *J. Cogn. Neurosci.*, 3: 72C86, 1991. 3

[80] T. Vetter. Synthesis of Novel Views From a Single Face. *International Journal of Computer Vision*, 28(2):103–116, 1998. 55

[81] P. Viola and M. Jones. Rapid Object Detection Using A Boosted Cascade of Simple Features. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001. ix, 84, 85

[82] F. Wallhoff. Facial Expressions and Emotion Database, 2006. URL http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html. 54

[83] G. Wallis and E. Rolls. A Model of Invariant Object Recognition in The Visual System. *Progress in Neurobiology*, 51:167–194, 1997. 8

[84] L. Wiskott, J. Fellous, N. Kruger, and C. v.d.Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7): 775–779, 1997. 12

[85] C. W.Zheng, X.Zhou and L.Zhao. Facial Expression Recognition Using Kernel Canonical Correlation Analysis (kcca). *IEEE Trans. Neural Networks*, 17(1):233–238, 2006. 37, 59, 80

[86] C. Xiang, X. A. Fan, and T. H. Lee. Face Recognition Using Recursive Fisher Linear Discriminant. *IEEE Transactions on Image Processing*, 15(8):2097–2105, Aug 2006. 44, 54

[87] YALE. The Yale Face Database. URL http://cyc.yale.edu/projects/yalefaces/yalefaces.html. 54

[88] M. Yeasin, B. Bullot, and R. Sharma. From Facial Expression to Level of Interest: A Spatio-temporal Approach. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 922–927, 2004. 81

[89] Y.Horikawa. Facial Expression Recognition Using KCCA with Combining Correlation Kernels and Kansei Information. In *Proc. Fifth International Conference on Computational Science and Applications*, pages 489–495, Perugia, Italy, 2008. 80

[90] L. Zhang, S. Li, Z. Qu, and X. Huang. Boosting Local Feature Based Classifiers for Face Recognition. In *IEEE Conf. Computer Vision and Pattern Recognition Worshop on Face Processing in Video*, Washington, DC, 2004. 12

[91] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between Geometry-based and Gabor-wavelets-based Facial Expression Recognition Using Multi-layer Perceptron. In *Proc. of third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 454–459, 1998. 11, 20, 37, 73

[92] G. Zhao and M. Pietikäinen. Dynamic Texture Recognition Using Local Binary Patterns with An Application to Facial Expressions. *IEEE. Trans. Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. 59, 81

[93] J. Zou, Q. Ji, and G. Nagy. A Comparative Study of Local Matching Approach for Face Recognition. *IEEE Transactions on Image Processing*, 16(10):2617–2628, 2007. 12

Publication List

Journal Papers

1. W.F.Gu, Y.V.Venkatesh, C.Xiang, A Novel Application of Self-Organizing Network for Facial Expression Recognition from Radial Encoded Contours, *Soft Computing*, vol.14, no.2, pp.113-122, 2010.

2. W.F.Gu, Y.V.Venkatesh, C.Xiang, D.Huang, H.Lin, Facial Expression Recognition using Radial Encoding of Local Gabor Features and Classifier Synthesis, *Pattern Recognition*, accepted, 2011.

3. W.F.Gu, Y.V.Venkatesh, C.Xiang, Web-based Facial Expression Recognition System using Radial Encoded Gabor Features and Classifier Synthesis, submitted to *Pattern Analysis & Application*, 2011.

Conference Papers

1. W.F.Gu, C.Xiang, H.Lin, Modified HMAX Models for Facial Expression Recognition, in *Proc. of the 7th International Conference on Control and Automation*, pp.1509-1514, New Zealand, 2009.

2. W.F.Gu, Y.V.Venkatesh, C.Xiang, Composite Orthonormal Basis for Person-Independent Facial Expression Recognition, In *Proc. of International Conference on Industrial Engineering and Engineering Management 2010*, pp.1942-1946, Macau, 2010.