A WEIGHTED REGIONAL VOTING BASED ENSEMBLE OF MULTIPLE CLASSIFIERS FOR FACE RECOGNITION

by

Jing Cheng

B.I.S., Tianjin University, 2007
B.A., Tianjin University, 2007

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
MATHEMATICAL, COMPUTER, AND PHYSICAL SCIENCES
(COMPUTER SCIENCE)

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

December 2011

Your file Votre référence
ISBN: 978-0-494-87531-5

Our file Notre référence
ISBN: 978-0-494-87531-5

Table of Contents

List of Tables

List of Figures

Abstract

Face recognition is heavily studied for its wide range of application in areas such as information security, law enforcement, surveillance of the environment, entertainment, smart cards, etc. Competing techniques have been proposed in computer vision conferences and journals, no algorithm has emerged as superior in all cases over the last decade. In this work, we developed a framework which can embed all available algorithms and achieve better results in all cases over the algorithms that we have embedded, without great sacrifice in time complexity.

We build on the success of a recently raised concept - Regional Voting. The new system adds weights to different regions of the human face. Different methods of cooperation among algorithms are also proposed. Extensive experiments, carried out on benchmark face databases, show the proposed system's joint contribution from multiple algorithms is faster and more accurate than Regional Voting in every case.

Acknowledgements

First of all. I would like to thank my friends for their putting up with my impatience working on this thesis even as I interrupted their work in a consistent manner without strong reason.

I would like to thank Heinrich Butow, our computer system support person. I would also like to thank Dr. Youqin Wang, our HPC Senior Lab Instructor. Extensive experiments were carried out in an optimized way with the her help.

I would like to thank my family, for all of the encouraging (either spiritual or financial) support. Thanks to my humble and open minded parents. I gradually became both spiritually and financially independent.

Most importantly, I would like to thank my supervisor Dr. Liang Chen for his guidance on my behalf, in both life and study. Lastly, I would like to thank professors of the department and my committee members: Dr. Desanka Polajnar and Dr. Jueyi Sui for their encouragement and guidance in this and everything else.

Chapter 1

Introduction

Biometrics is tagged as one of the 'top ten emerging technologies that will change the world' in 2001 by the MIT Technology Review [Heyer, 2008]. With potential to stamp out forgery and theft possible with other methods, it is a good fit for security authentication problems.

The face stands out as a biometric compared to several other popular physiological or behavioral characteristics for the purpose of distinguishing between individuals:

1. The fingerprint can be defected by chemical contaminations;

2. The iris is not workable with diabetic victims;

3. The retina is not convenient during data collection for operational biometric systems;

4. The voice changes when a person's throat is infected and is sensitive to noise.

As a biometric, face recognition uses faces to identify a person. The history of face recognition, began with excessive optimism followed by scepticism. After a continuous effort to overcome the exposed limitations, face recognition has become 'one of the most successful applications of image analysis and understanding [Zhao et al., 2003].'

1.1 Motivation

Face recognition is broadly defined as assigning identity to one or more input face images with a registered identity in the database [Ijiri and Sakuragi, 2006]. Its usage is generally divided into: 1) identification and 2) verification. Identification is a "one-to-many" search to answer questions like "Who is he? " and is more common in surveillance. It is a closed set problem assuming that the input face image corresponds to an identity stored in the system. It is also referred to as a forced-choice experiment in the psychology literature [Moon and Phillips, 2001]. Verification is a "one-to-one" search to answer question like "Is he Eric?" Given a probe image and a claimed identity, a decision, either 'accept' or 'reject' has to be made. This is more frequently used for security access. The latter is an open set problem as there may be no corresponding identities stored in the system. This research deals with the "one-to-many" problem, but the system is believed to perform equally well on the "one-to-one" problem as well.

A face recognition system automatically identifies a human face from database images. It is challenging as it needs to account for all possible appearance variations caused by change in illumination, facial features, occlusions, etc [P. Latha and Annadurai, 2009]. It has drawn a huge surge of attention due to its commercial potential and cultural significance. The need for face recognition can be found in smart environment [Pentland and Choudhury, 2000], entertainment, smart cards, information security [Ijiri and Sakuragi, 2006], law enforcement and surveillance systems [Zhao et al., 2003].

Over the past semi-century, many approaches have been developed. All approaches require comparison of an input face image with face images labelled with known identities stored in a database to claim a match. There are two types of solutions for the comparison problem: the first is a step-by-step based decision-making process, and the other is a 'learning mechanism' based decision-making process such as a neural network. A 'step-by-step' system carries out executions previously designed and set. It does not interact with an environment or get modified based on the result accumulated up to that point. A learning mechanism is a more dynamic system, presumably using what it has learned. It identifies the information needed for its ongoing performance. Most face recognition systems belong to the former category. "A step-by-step" system is adopted in this thesis.

1.2 Major Contribution

In this thesis, I studied strategies for adopting a two-layer voting scheme for face recognition. A quick summary of the major contribution of this thesis is as follows:

1. Based on the regular face recognition procedure, I constructed a new face recognition system adopting a two-layer voting scheme. Five top face recognition algorithms over the last decade have been employed. They are:

Prominent traditional approaches include Principle Component Analysis (PCA) and Fisherface which have profound effects.
Newly developed algorithms published in leading journals and conferences: Spectral Regression Dimension Analysis (SRDA) [Cai Deng and Jiawei, 2008], Spatially Smooth Version of Linear Discriminant Analysis (S-LDA) [Cai D. and Huang, 2007] and Spatially Smooth Version of Locality Preserving Projection (S-LPP) [He Xiaofei and Hongjiang, 2005].

2. The system is not specific for embedding the algorithms mentioned above, but also available for some other-class approaches, local approaches such as Local Binary Patterns (LBP) [Timo Ahonen and Pietikainen, 2004] for instance.

3. In the newly proposed system, weight takes a key role. I proposed three methods for generating weights.

4. Extensive experiments, carried out on benchmark face databases, show the proposed system is faster and holds a lead in every case over an already proven system-Regional Voting which has been shown to be very stable in the face of a noisy environment. In a lot of cases, the weighting algorithms and ensemble among different embedded approaches proposed here reduce the error recognition rate by more than half. The same promising results on experiments of datasets with small number of images per person in the gallery images deserves emphasis as it belong to a especially sticky problem in the face recognition area: the SSS (small-sample-size) problem which will be further addressed in Chapter 3.

We have only just scratched the surface with this introduction for the time being. The thesis is broken into "chunks" designed to fill different needs. The following does not cover anything in depth, but instead gives a high-level overview of how the thesis is structured. The rest of the thesis is organized as follows: Chapter 2 traces the pertinent literature on face recognition. Chapter 3 elaborates the design of the proposed system. Chapter 4 demonstrates the performance of the system by results from the experiments followed by a result analysis in chapter 5. Chapter 6 summarizes the system presented and provides suggestions for future work.

Chapter 2

Literature Review

To distinguish a person from hundreds of others by just a mere glance despite variations in viewpoint, lighting, emotional expression and hairstyle, face recognition is one of the most amazing features of the visual system [McKone E. and N., 2009]. Previous efforts on prosopagnosia (a symptom that people fail to recognize human faces) have unveiled some properties of how human brain mechanisms function [Tranel and Damasio, 1985]. In 1991, against the conventional opinion for over the past 30 years that face recognition develops very slowly throughout infancy, childhood and adolescence [McKone E. and N., 2009], a two-process theory of infant face recognition based on experiments on infants supported the conclusion that infants are born being aware of information and structure of faces. It further brought forth two terminologies: CONSPEC and CONLEARN to reveal how infants foster their ability in face recognition gradually[Morton and Johnson, 1991]. CONSPEC guides the preference for facelike patterns for newborn infants. CONLEARN is responsible for learning visual characteristic. A review of the effect of inversion upon face recognition claimed that face recognition is 'special' [Valentine, 1988]. This phenomenen was further validated in 1996 by evidence from neuropsychology [Farah, 1996]. In 1997, a cross-species study on face recognition in primates: monkey and adult humans, showed that both species have novel preference for their own species [Pascalis and Bachevalier, 1998]. In 2001, a meta-analytic review claimed that own-race faces are better remembered compared to other faces [Meissner and Brigham, 2001]. Over the last two decades in neuroethology, cognitive and neural mechanisms have been looked into with focus on their potential roles in enabling humans to recognize faces [McKone E. and N., 2009].

Similar to the behavior of a human brain that associates memory to distinguish between faces, face recognition tries to retrieve from a gallery of images labeled with known identities and assign an identity to a given probe image. It is easy to describe but hard to implement. Even though humans have always had the innate ability to recognize faces, automatic face recognition systems are still in an early stage either due to the lack of awareness of how human cognition works, or the infeasibility at computationally modeling billions of neurons. So, different approaches have been proposed for the face recognition problem. The rest of this chapter is divided into seven sections. The first section starts with a review of the procedures of a face recognition system, followed by an overview of solutions in the second section. Then, the most popular solution, the holistic approach, is elaborated, illustrating five important algorithms within this family. The last section is a review of a recently proposed technique which has achieved great success: regional voting.

2.1 Procedure of Face Recognition Systems

The procedure can be broken into six segments, to be handled regardless of the specific method used [Zhao et al., 2003]:

1. : Capture image
2. : Face location (detection) in image
3. : Face image pre-processing
4. : Feature extraction
5. : Template comparison
6. : Match declaration

Figure 2.1 shows a more detailed processing flow of a face recognition system.

First, the image is assumed as a structured collection of pixels. The acquisition can be accomplished either by scanning existing photographs or shooting live pictures of a person with a camera. Usually several samples have to be taken for a bigger possibility to be matched. Video is also considered as a source, as it consists of a sequence of still images. Detecting the face in an image alone is a hard task. Once the face images have been targeted, pre-processing starts the refinements on the images. Figure 2.2 shows the detailed steps in this segment.

Pre-Processing includes the following four steps as shown in order: (1) finding the location of the pupil, (2) rotating to have pupils aligned, (3) scaled and (4)

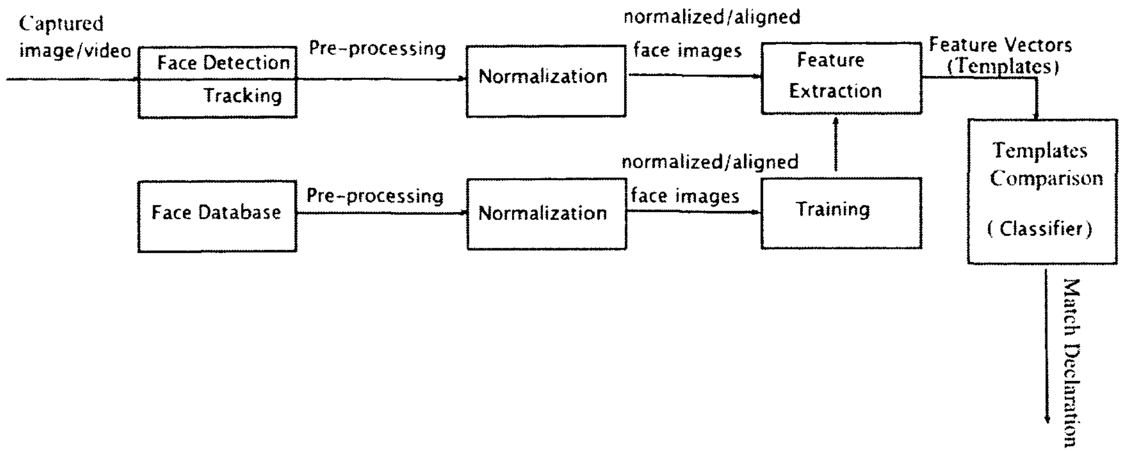


Figure 2.1: Face Recognition Processing Flow

cropped. The "original image" comes from the Olivetti Research Laboratory (ORL) database[1]. The following three images having pupils darkened show the intermediate stages during the process. At last, after cropping, images finally become the objects which are going to be classified by face recognition systems. To stay focused on the proposed system, we assume that the four steps have already taken place. For an overview of each step, please see Gonzalez et al [Gonzalez et al., 2009]. The algorithms on how Pre-processing plays around the eyes can be found in P. Wang et al [P. Wang and Wayman, 2005].

(a) Original Image (b) Pupil Locating (c) Rotating (d) Scaling (e) Cropping

Figure 2.2: Pre-Processing

After pre-processing, images are cropped to a smaller size with pupils aligned. A face image becomes an $h \times w$ matrix of pixel values having each of them standing for a colour value (h and w refers to the height and width of the image respectively). The

[1] AT & T (ORL) Database: `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`

pixel value here is a single real number representing the intensity of gray particularly. Grayscale images are employed in this thesis. More recently, face recognition systems use 3D images to analyze skin or skull geometry. But some 3D face recognition systems are expensive such as the ones using laser to capture images. Tao et al [Tao Q and Weber, 2007] constructed a multimodal or fusion approach for facial verification which shows that a combination of both 2D and 3D information significantly improves performance. Generally, 2D images are adopted by most current face recognition systems due to a better cost/benefit trade off.

Then comes the key segment of a face recognition system: feature extraction. For holistic approaches, Dimension reduction algorithms will be used here. After this, the full length face vectors are projected to a lower dimensional subspace. Each face image is represented by a much shorter vector with redundant information removed. The output of this segment is a set of feature vectors in a lower dimensional space. They are referred to as templates representing an enrollee's face. In the next (fifth) segment, the generated templates are compared with the enrolled templates stored in the database. Finally, the classifier yields a score on how close they are matched in order to identify the enrollee.

2.2 Overview of Approaches

Despite of the fact that there are methods for commercial use which are proprietary by the vendors, broadly speaking, all the methods fall into the following four categories: local approaches, holistic approaches, neural networks and automatic face processing [Das, 2011]. All approaches share some traits and differ from each other with their own specialties in the meantime. For instance: all are best suited with front-on images and a well-lit environment.

Local approaches attempt to extract specific features from different regions of face images. They are advantageous over other facial recognition system in their robustness against variance in appearance or angles a face is presented towards input sensors(2D video or digital camera).

Holistic approaches incorporate the entire face image. A grayscale image of h pixels tall and w pixels wide is represented as a slim vector of length $h \times w$. Each face image (slim vector) can be interpreted as a point in a high dimensional space

$\mathrm{R}^{h \times w}$. Each pixel in an image can be taken as a coordinate in $\mathrm{R}^{h \times w}$. Natural images contain clouded, unclear and redundant data[Ruderman, 1994]. Again, the redundancy problem for face images particularly is deteriorated by normalization with respect to scaling, translation and rotation [Penev and Atick, 1996, Zhao W. and J., 1999]. Thus, dimension reduction of the long thin vectors to get compressed data with redundancies removed is the core of this set of approaches.

Given gallery images: $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$, the projection matrix P is trained on the gallery images by solving a generalized matrix problem:

$$GP = B \tag{2.1}$$

P is a linear projection function matrix ($\kappa \times hw$). After projection: $G^{n \times hw} \xrightarrow{P} B^{n \times \kappa}$, images with highlighted features are kept in a lower dimensional subspace ($\kappa \ll hw$). The same dimension reduction technique is implemented on probe images Y. Given a probe image $y \in Y$, we compute $y_P = yP'$. Now with both probe images and gallery images in a lower dimensional subspace, there is a set of measurements to compare the similarities between feature vectors: L_1 distance, L_2 distance, Mahalanobis distance, angle between feature vectors, etc. L_2 (Euclidean) distance is used in this thesis. Taking Euclidean distance as a reference for closeness, the classification can be ranked. Finally, some classification method is implemented to make the selection and the nearest neighbourhood is chosen so that the best match between y_P and the rows of B can be found (each row in B corresponds to the identity of one of the images G) and the closest identity is assigned to the probe image. Five holistic approaches that will be embedded in the proposed system are elaborated in detail in the following sections as examples illustrating different techniques the projection matrix, P, is adopting.

Neural network approaches have addressed several issues: gender classification, face recognition and facial expression classification [R. Chellappa and Sirohey, 1995]. The earliest face recall application using neural networks (NN) is reported in Kohonen's associative map [Kohonen, 1988]. A device named WISARD for face recognition, expression analysis and face recognition using a single layer adaptive NN is reported by Stonham [Stonham, 1986]. Classification is achieved by determining the classifier which replies strongest to the input image. Later systems based on dynamic link architecture (DLA) have been proposed [M. Lades and Wurtz, 1993, Ellis, 1986].

There is a dynamic variable (J) between two neurons (i,j). J_{ij} will be increased or decreased if there are positive or negative connections between neurons respectively and finally return to a resting state; thus, the weights that are associated with certain features can be modified. Neural network mapping attempts to utilize as many features as possible to ascertain a match between the enrolment and references in the database [Heyer, 2008].

Automatic face recognition systems acquire matches using distance or dependent variables of distance (distance ratios for instance) between salient features of the face. It is almost the simplest approach; thus, it does not have parameters to be tuned to different environments: dim light for instance [Heyer, 2008].

The latter two categories are listed above due to the consideration of integrity. The former two are more competitive and thus have been frequently referred to regarding solutions for the face recognition problem. A hybrid approach which combines both techniques is also categorized as an individual branch of solutions for face recognition problems.

2.3 Principal Component Analysis (PCA)

PCA, a valuable technique from applied linear algebra, selects a subspace preserving as much variation as possible [Ahmed and Rao, 1975]. The first breakthrough among the subspace approaches [Wang and Tang, 2003], it is used abundantly in analysis from neuroscience to computer vision for its simple non-parametric properties extracting relevant information from large or confusing data [Shlens, 2005]. Turk and Pentland first used it in face recognition in 1991 and named it as eigenface which is a reflection of the nature of the calculation [Turk and Pentland, 1991]. By reconstructing face images from lower dimensional subspace accounting for as much variance as possible [Ahmed and Rao, 1975], PCA has been one of the driving forces behind a broad spectrum of studies, including face detection, recognition [Zhao et al., 2003] and sex classification [Ahmed and Rao, 1975].

First, we create a mean centred matrix G_m using gallery images: $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$

and n is the total number of gallery images.

$$G_m = \begin{bmatrix} \mathbf{g}_1 - \overline{\mathbf{g}} \\ \mathbf{g}_2 - \overline{\mathbf{g}} \\ \mathbf{g}_3 - \overline{\mathbf{g}} \\ \vdots \\ \mathbf{g}_n - \overline{\mathbf{g}} \end{bmatrix} \tag{2.2}$$

where $\overline{\mathbf{g}} = \sum_{i=1}^{n} \mathbf{g}_i / n$.

Then we arrive at a definition for the covariance matrix C_G.

$$C_G \equiv \frac{1}{n-1} G_m {G_m}^T \tag{2.3}$$

Thus the diagonal terms of C_G denote the variance of particular measurement types: the pixel values standing for the intensity of gray in different dimensions in this case[2]. Our goal is to maximize the variance and minimize the redundancy which is measured by covariance of measurement types; more specifically, the off-diagonal terms of C_G. An ideal covariance matrix is an "optimized" matrix having all off-diagonal terms "0". Diagonalization can be done by finding some orthonormal matrix P where $G_m = PX$. P is constituted of the principal components of X. By substituting G_m in the above equation 2.3, we get:

$$C_G = \frac{1}{n-1} P(XX^T)P^T \tag{2.4}$$

In the above equation, we get a symmetric matrix XX^T. This can be further diagonalized by an orthogonal matrix of its eigenvectors:

$$XX^T = EDE^T \tag{2.5}$$

where D is a diagonal matrix and E is a matrix of eigenvectors of XX^T arranged in columns. In order to avoid massive calculation, we select eigenvectors of XX^T as rows of matrix P. Thus, $P \equiv E^T$. Substituting into equation 2.5, C_G will finish evaluating by:

[2]Note: $n - 1$ in the equation is the proper normalization for an unbiased estimator.

$$
\begin{aligned}
C_G &= \frac{1}{n-1} P X X_T P^T \\
&= \frac{1}{n-1} P (P^T D P) P^T \\
&= \frac{1}{n-1} (P P^T) D (P P^T) \\
&= \frac{1}{n-1} (P P^{-1}) D (P P^{-1}) \\
&= \frac{1}{n-1} D
\end{aligned}
$$

Eventually, the subspace (eigenspace) in this case is spanned by k eigenfaces with the largest eigenvalues. PCA has become a de facto benchmark algorithm due to its ease of implementation and its reasonable performance levels [Phillips P J and S, 1997]. The non-parametric property can be also viewed as a weakness for there are few tuning opportunities. Some prior non-linear transformations (sometimes termed as kernel transformations) are introduced to PCA and have been proposed to incorporate selected parameters with known priorities. Common kernel transformations include Fourier and Gaussian transformations and the entire parametric algorithm is named kernel PCA. PCA constrains the data to be Gaussian distributed. The "Gaussian distributed" constraint has only recently been solved via ICA (Independent Component Analysis) [Shlens, 2005]. Though ICA is a form of nonlinear optimization which is powerful for solving a new class of problem, exponentially distributed data for instance, it is difficult to calculate in practice.

2.4 Fisherface

Belhumeur proposed Fisherface in 1997. It aims to maximally discriminate intra-class and inter-class and is claimed to be strong under large variations in illumination and facial expressions [Belhumeur. P and D, 1997]. Lighting variability includes intensity, direction and light sources. Variations between images of the same person caused by lighting variabilities are larger than those due to different identities [Y. Moses and Ullman, 1994]. Although PCA is optimal for reconstruction from a lower dimensional basis, Fisherface advances over it from a discrimination point of view. Fisherface is a derivative of Fisher's Linear Discriminant (FLD). It is a "classical" technique in pattern recognition [Duda and Hart, 1973] and was first developed by Robert Fisher in 1936 for taxonomic classification [Fisher, 1936].

In general, Fisherface uses a class specific linear method: Fisher's Linear Discriminant (FLD) [Fisher, 1936] for dimensionality reduction and a simple classifier in the low dimensional space of feature vectors. More formally, in order to "shape" the scatter, we use S_W and S_B denoting the within class scatter matrix and the between class scatter matrix separately. More formally, we assume that the classification is among n different identities. Let G, G_m and n be defined as in the eigenface approach. Let ψ stand for the the number of different subjects to be classified. They are calculated by the following equations.

$$S_W = \sum_{i=1}^{\psi} P_{r(\omega_i)} S_i \tag{2.6}$$

$P_{r(\omega_i)}$ is the prior class probability. In practice, with the assumption of equal priors, it is usually replaced by $\frac{1}{\psi}$. S_i: the covariance of the images of class ω_i is the sample vectors $\mathbf{g}$ of around its mean μ_i:

$$S_i = E[(g(\omega) - \mu_i)(g(\omega) - \mu_i)^T | \omega = \omega_i] \tag{2.7}$$

$$S_B = \frac{1}{\psi} \sum_{i=1}^{\psi} (\mu_i - \overline{\mathbf{g}}) \cdot (\mu_i - \overline{\mathbf{g}})' \tag{2.8}$$

where n and μ_i are defined the same as in equation 2.7, and $\overline{\mathbf{g}}$ is the overall mean for gallery data. The projection matrix P we are looking for this time is oriented to satisfy the following equation:

$$P = \arg\max_P \frac{|P^T S_B P|}{|P^T S_W P|} = [p_1 p_2 \cdots p_m] \tag{2.9}$$

where $\{p_i | i = 1, 2, \cdots, m\}$ are generalized eigenvectors corresponding to m largest eigenvalues $\{\lambda_i | i = 1, 2, \cdots, m\}$ of S_B and S_W:

$$S_B \mathbf{p}_i = \lambda_i S_W \mathbf{p}_i, i = 1, 2, \cdots, m \tag{2.10}$$

where m is upper bounded to $\psi - 1$, for the number of nonzero generalized eigenvalues is at most $\psi - 1$ [Duda and Hart, 1973].

This works pretty well for general finding projection problems. While in the face recognition problem, the matrix S_W is probably singular for the number of classes to be classified, n is far less than the number of dimensions, hw. This threatens the denominator in equation 2.9. $|P^T S_W P|$ is exposed to the risk of becoming zero as no such projection P can be found. Belhumeur et al [Belhumeur. P and D, 1997] overcame the complication of a singular S_W by a method called Fisherface. PCA is first used to reduce the dimension of the feature space to $n - \psi$, where an alternative criterion is proposed to 2.9:

$$P_{opt}^T = P_{fld}^T P_{pca}^T \tag{2.11}$$

then Fisher's Linear Discriminant, defined in 2.9, is used to reduce the dimension to $\psi - 1$.

$$P_{pca} = \arg\max_P \{P_{pca}^T C_G P_{pca}\} \tag{2.12}$$

where C_G is computed by equation 2.3.

$$P_{fld} = \arg\max_P \frac{|P_{fld}^T P_{pca}^T S_B P_{pca} P_{fld}|}{|P_{fld}^T P_{pca}^T S_W P_{pca} P_{fld}|} \tag{2.13}$$

Note that the optimization for P_{pca} is performed over $hw \times (n - \psi)$ matrices. hw is the number of pixels in an image of height h and width w. Constrained to the condition that the data for each class is approximately Gaussian distributed [Yan et al., 2005] again, Fisherface encodes information in a linear separable space unlike the orthogonal linear space the way PCA encodes. Extensive experiments have been carried out comparing eigenfaces versus Fisherfaces [Belhumeur. P and D, 1997]. The results have demonstrated Fisherface's superiority to PCA in handling variations in lighting and expression.

2.5 Spectral Regression

The two most popular holistic face recognition methods, unsupervised PCA and supervised Fisherface, deteriorate rapidly when there are large variations in viewpoint, illumination or facial expression [Juwei Lu and Li, 2006]. Linear Discriminant Analysis (LDA), which the Fisherface method is based on, has been widely used in fields such as machine learning, data mining, information retrieval, and pattern recognition

[Cai Deng and Jiawei, 2008]. The projection functions of LDA extract features preserving class separability by maximizing the between class covariance and minimizing the within-class covariance simultaneously. However, the computation of LDA involves dense calculations on matrices. For instance, eigen - decomposition, which can be computationally expensive both in time and memory, is involved twice when the scatter matrix is singular. Thus, utilization of LDA on large scale high dimensional data is infeasible. Spectral Regression Dimension Analysis (SRDA) [Cai et al., 2007b, Cai Deng and Jiawei, 2008], built on the framework of Graph Embedding [Yan et al., 2005], is both a framework and an algorithm. It reduced the computational complexity of the dimension reduction techniques (a more efficient LDA is introduced in [Yan et al., 2005]).

Suppose there are n $h \times w$ face images. A symmetric $n \times n$ matrix can be constructed having the weight of the edge joining vertices i and j W_{ij} as entries. The collection of vertices can be denoted in a vector representation: $\{g_i\}_{i=1}^{n} \subset \mathrm{R}^{h\times w}$ and $G = [g_1, \cdots, g_n]$. This graph reduction model is oriented to represent the vertices of the graph as a vector with dimension lower than hw. The reduction matrix B (an $n \times \kappa$ matrix) of G to a lower dimensionality κ:

$$B = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \vdots \\ \mathbf{b}_n \end{bmatrix} \tag{2.14}$$

should minimize the following:

$$\sum_{i,j} (b_i - b_j)^2 W_{ij} \tag{2.15}$$

That is, a heavy penalty will be incurred if neighboring vertices i and j are far apart. B^* (optimal B) can be found by solving equation

$$\mathbf{b} = \arg\max_{\mathbf{b}} \frac{|\mathbf{b}' W \mathbf{b}|}{|\mathbf{b}' D \mathbf{b}|} \tag{2.16}$$

where D is a diagonal matrix whose entries are column (or row, since W is symmetric) sums of W, $D_{ii} = \sum_j W_{ji}$) and the solutions to this equation are the eigenvector solutions to

$$Wb = \lambda Db \tag{2.17}$$

The optimization given in equation (2.16) can be reformulated via a linear translation from G to B: $PG = B$ as: p = $\arg\max_{\mathbf{p}} \frac{|\mathbf{p}'G'WG\mathbf{p}|}{|\mathbf{p}'G'DG\mathbf{p}|}$

Finding eigenvalues is computationally exhaustive and, furthermore, can't be done if GDG' is non-singular when the number of images is far less than the number of features in the image. To get around this issue, Cai et al [Cai et al., 2007b] show that p will be a solution to equation 2.5 if $\mathbf{b} = \mathbf{p}G$ satisfies equation 2.17.

In a more efficient LDA [Yan et al., 2005], W is determined as:

$$W_{ij} = \begin{cases} \frac{1}{n_c} & \text{if vertex } i \text{ and vertex } j \text{ both belong to the } c\text{th class} \\ 0 & \text{otherwise} \end{cases} \tag{2.18}$$

where n_c is the number of samples from class c. While this leads to a useless eigenvector: $\mathbf{b}_i = [1, 1, 1, \ldots, 1]$ with associated eigenvalue 1 in the orthogonal basis, the Gram Schmidt algorithm is applied here to get around this problem, deriving a $\psi - 1$ dimensional basis.

After having found the basis, B, via Gram Schmidt process, the matrix P in $B = PG$ can be approximated by:

$$\mathbf{p} = \arg\min_{\mathbf{p}} \sum_{j=1}^{n} (\mathbf{p}\mathbf{g}_j - \mathbf{b}_{ij}) \tag{2.19}$$

where $\mathbf{b}_{ij}$ is the jth element of vector $\mathbf{b}_i$ and g_j is the jth vector of G. Least squares approximation technique such as LSQR [Paige and Saunders, 1982] are popular solutions for this.

The number of computations performed is drastically reduced and more complex methodologies could be embedded in this framework. Experimental results [Cai et al., 2007b,a] have shown that SRDA (spectral regression dimension analysis) outperforms LDA.

2.6 Spatially Smooth Subspace Learning

Holistic appearance-based face recognition algorithms such as the ones introduced above: Principal Componant Analysis (PCA), Linear Discriminant Analysis (LDA) which Fisherface is based on have attracted considerable interests in recent years [Cai D. and Huang, 2007]. Many subspace space learning algorithms such as Locality Preserving Projection (LPP) [He and Niyogi, 2003], Neighborhood Preserving Embedding (NPE) [He et al., 2005], Marginal Fisher Analysis (MFA) [Yan et al., 2005] and Local Discriminant Embedding (LDE) [H.-T. Chen and Liu, 2005], have been proposed to estimate the geometrical and topological properties of the submanifold from random points ("scattered data") lying on this unknown submanifold. What if the face image lying on a nonlinear submanifold [He Xiaofei and Hongjiang, 2005] is missing? Besides, all the above methods consider an image of size $h \times w$ a vector (point) in $R^{h \times w}$ and the pixel values in the vectors independent. While the intrinsic representation of an image in a plane is a matrix, or 2-order tensor, the spatial relationship of the pixels is lost in this case. A "Spatially Smooth Subspace" proposed for Face Recognition using a Laplacian penalty to constrain the coefficients to be spatially smooth fixed the above problems. This model can feed all the existing subspace learning methods [Cai D. and Huang, 2007].

The remainder of this section is structured as follows: Subsection 2.6.1 provides a brief review of the two subspace learning algorithms: LPP and LDA. Susection 2.6.2 introduces their tensor extension. Subsection 2.6.3 details the Spatially Smooth Subspace Learning (SSSL) model followed by experimental results and concluding remarks.

2.6.1 Locality Preserving Projections (LPP) and Linear Discriminant Analysis (LDA)

Other approaches that use the graph embedding model outlined above in section 2.5 include: Locality Preserving Projections[He and Niyogi, 2003] (LPP), Linear Discriminant Analysis (LDA), NPE, MFA and LDE with different choices of W and D. The choice for LDA has been introduced in the previous section so that the choice of W for LPP (another algorithm embedded in the proposed system) is briefly listed below.

Again, the LPP graph is constructed by placing an edge between two vertices if they are "close." Two criteria for determining "closeness" are suggested: k-nearest neighbours and ϵ distance [He Xiaofei and Hongjiang. 2005].

$$W_{ij} = \begin{cases} \exp\left(\frac{\|\mathbf{g}_i - \mathbf{g}_j\|^2}{\eta}\right) & \text{if } \|\mathbf{g}_i - \mathbf{g}_j\|^2 < \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{2.20}$$

or

$$W_{ij} = \begin{cases} \exp\left(\frac{\|\mathbf{g}_i - \mathbf{g}_j\|^2}{\eta}\right) & \text{if } j \text{ is among the } k \text{ nearest neighbours of } i \\ 0 & \text{otherwise} \end{cases} \tag{2.21}$$

where $\eta \in \mathbb{R}$ is a tuning parameter.

2.6.2 Tensor Extensions

The relationship between nearby pixels of the face images might be important for finding a projection. In order to keep the spatial relationship of the pixels, the tensor-based approaches operate directly on the matrix representation of face images.

Let $\{u_k\}_{k=1}^{h}$, $\{v_l\}_{l=1}^{w}$ be the orthonormal basis of R^h and R^w separately. It is shown that $u_i \otimes v_j$ forms a basis of the tensor space $\mathrm{R}^h \otimes \mathrm{R}^w$ [Lee, 2002]. More specifically, the projection of $T \in \mathrm{R}^h \otimes \mathrm{R}^w$ on the basis $u_i \otimes v_j$ can be computed as their inner product:

$$\langle T, u_i \otimes v_j \rangle = \langle T, u_i v_j^T \rangle = u_i^T T v_j \tag{2.22}$$

Unlike the ordinary vector-based approaches which are linear, i.e. $y_i = a^T g_i$ where $g_i \in \mathrm{R}^{hw}$ is the vector representation of the i-th image and y_i is the lower dimensional vector after projection by dimension reduction vector (basis vector) a. The tensor-based approaches are multilinear, i.e. $y_i = u^T T_i v$ where $T_i \in \mathrm{R}^h \otimes \mathrm{R}^w$ is the matrix representation of the i-th image and $\aleph = h \times w$. In a tensor basis uv^T, the degrees of freedom are down to $h + w$. In fact, with the following constraint, the tensor-based approach becomes, a special case of vector-based approaches:

$$a_{i+h(j-1)} = u_i v_j \tag{2.23}$$

2.6.3 Spatially Smooth Subspace Learning

Intuitively, the element pixel values in basis function would be similar if they are spatially close. Intending to suggest the spatial smoothness of the basis function by the spatial correlation of pixels in a face image, a 2-D discretized Laplacian penalized function is introduced to measure the smoothness of the basis vectors of the face space which is the core of the Spatially Smooth Subspace Learning (SSSL) approach.

Let y_i stand for a vector in R^{hw} representing an $h \times w$ as described above, $a_i \in \mathrm{R}^{hw}$ be the bases vectors (projection functions) and $\aleph = h \times w$. The region of interest, the face image, which is a two-dimensional rectangle face image is denoted by Ω or $[0,1]^2$ for notational convenience. Let $\zeta = (h_1, h_2)$ where $h_1 = 1/h$ and $h_2 = 1/w$. Ω_ζ is constituted of two-dimensional vectors $t_i = (t_{i_1}, t_{i_2})$ with $t_{i_1} = (i_1 - 0.5) \cdot h$ and $t_{i_2} = (i_2 - 0.5) \cdot w$ for $1 \leq i_1 \leq h$ and $1 \leq i_2 \leq w$. The total number of grid points in this lattice is $\aleph$. Let D_j be an $h \times h$ or $w \times w$ matrix that yields a discrete approximation to $\partial^2/\partial t_j^2$. Thus if $u = [u(t_1), \cdots, u(t_{n_j})]$ (where $1 \leq j \leq 2$ and $n_1 = h, n_2 = w$) is an n_j dimensional vector which is a discretized version of a function $u(t)$, then D_j has the property that:

$$[D_j u]_i \approx \frac{\partial^2 u(t_i)}{\partial t^2} \tag{2.24}$$

for $i = 1, \cdots, n_j$. Many possible choices are available for D_j [B. L. Buzbee and Nielson, 1970] and the following is the modified Neuman discretization [O'Sullivan, 1991]:

$$D_j = \frac{1}{h_j^2} \begin{bmatrix} -1 & 1 & & & & & 0 \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & . & . & . & & & \\ & & 1 & -2 & 1 & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -1 & \end{bmatrix} \tag{2.25}$$

Given D_j where $1 \leq j \leq 2$, a discrete approximation for two-dimensional Laplacian L is the $\aleph \times \aleph$ matrix:

$$\Delta = D_1 \otimes I_2 + I_1 \otimes D_2 \tag{2.26}$$

where I_j is the $n_j \times n_j$ identity matrix for $j = 1, 2$ and $\otimes$ is the Kronecker product [Horn and Johnson, 1991]. For a $n_1 \times n_2$ dimensional vector γ, to check its smoothness on the $n_1 \times n_2$ lattice, we can check that $|| \Delta \cdot \gamma ||^2$ is proportional to the sum of the squared differences between nearby grid points of γ with its matrix form. Given a pre-defined graph structure with weight matrix W, the SSSL approach is oriented for maximizing the following:

$$\frac{\gamma^T GWG^T \gamma}{(1-\alpha)\gamma^T GDG^T \gamma + \alpha \Xi(\gamma)}, \tag{2.27}$$

where $0 \leq \alpha \leq 1$ controls the smoothness of the estimator and Ξ is the discretized Laplacian regularization function:

$$\Xi(\gamma) = || \Delta \cdot \gamma ||^2 = \gamma^T \Delta^T \Delta \gamma \tag{2.28}$$

The γ_{opt} which maximizes the objective function 2.29 can be derived by the maximum eigenvalue solutions to the following general eigenvalue problem.

$$GWG^T \gamma = \lambda((1-\alpha)GDG^T + \alpha \Delta^T \Delta)\gamma \tag{2.29}$$

With the options of different W as described in subsection 2.6.1, spatially smooth version of LDA and LPP are hence derived. He et al [He Xiaofei and Hongjiang, 2005] showed a dramatic improvement of S-LPP over PCA and LDA for face recognition on the the Carnegie Melon University Pose, Illumination, and Expression database (CMU PIE)[3] and Yale[4] databases. S-LDA and S-LPP developed methods based on the SSSL model outperform the ordinary subspace learning algorithms and their tensor extensions. Zheng et al demonstrate SLPP showing a significant improvement over standard LPP [Z Zheng and Yang, 2007].

2.7 Regional Voting

Large variations in viewpoints, illumination or facial expressions always lead to a highly nonconvex and complex distribution of face images [Bichsel and Pentland, 1994]. Thus their success is limited to their linear nature [Juwei Lu and Li, 2006]. Either nonlinear models or a mixture of locally linear models could handle the nonconvex problem [Juwei Lu and Li, 2006]. Regional Voting, a framework proposed by Chen and Tokuda[Chen and Tokuda, 2010], is one solution for the latter case. It embeds all holistic algorithms. The image is broken into non-overlapping equally sized

[3]CMU PIE database: `http://www.ri.cmu.edu/research_project_detail.html?project_id=418&menu_id=261`.

[4]Yale Database `http://cvc.yale.edu/projects/yalefaces/yalefaces.html`.

windows. Holistic classification and national voting are applied inside each window (region). After all the windows are enumerated, record the result for each region. The majority winning windows has the last word.

This method is a complementary to holistic approaches which lack the knowledge of the spatial structure of the face and is more accommodating to 'noise'. Chen and Tokuda reconstruct face recognition as a 'stability' problem [Chen and Tokuda, 2010]. Under this concept, the probe image has undergone transformations caused by the environment: 'being obliterated' for instance. From this point of view, this system will still work as a whole even if some regions have been occluded. A deeper analysis of regional voting over national voting can be seen in Chen and Tokuda [Chen and Tokuda, 2005, 2003].

While since each window could work as a classifier, it naturally raises questions on the topic of classifier combination [Kittler and Matas, 1998]. Combining classifiers works best when they are different [Ali and Pazzani, 1995]. For example, two popular ensembles of classifiers employ majority voting, based on labels and label ranking respectively [Bagui and Pal, 1995], [T.K. Ho and Srihari, 1994]. The hope, when multiple algorithms are embedded into one framework, is they will complement one other and contribute jointly to the final decision making.

Chapter 3

Proposed Approach

Holistic face recognition approaches based on statistical learning, such as the ones based on LDA, often suffer from the SSS (small-sample-size) problem, where the dimensionality of the sample images far exceeds the number of training sample images available for each subject [Raudys and Jain, 1991], [Juwei Lu and Li, 2006]. Building on the success of Regional Voting, we present a system called Weighted Regional Voting Based Ensemble of Multiple Classifiers (WREC) for face recognition. This idea exploits the fact that face regions are of different significance when recognizing a face. This concept can be traced back as early as 1970's in "computer recognition of human faces" [Kanade, 1977]. Recent exploitations of the weight distribution can refer to "A Weighted Voting and Sequential Combination of Classifiers Scheme for Human Face Recognition [Xiaoyan Mu and Watta, 2005]", Local Binary Patterns by Timo Ahonen and others [Timo Ahonen and Pietikainen, 2004]. An automatic weighting evaluation is implemented in this thesis so that a more robust system having a larger capacity of bearing the variance of the face images is constructed. It is independent of human knowledge of the underlying structure of the face. For instance, images with exaggerated facial expressions in which symmetry has been violated is still tolerated by the system. The ensemble of multiple face recognition algorithms is motivated by the fact that different algorithms address different obstacles in face recognition. Though new algorithms have been added to the face recognition literature, none of them is able to integrate all the advantages into one. WREC provides an interface for different algorithms bringing their good attributes into full play. It has the following key features:

1. Images are partitioned into non-overlapping regions;
2. For each region, multiple holistic algorithms are employed;
3. A weight is associated with each holistic algorithm for each region;

4. The decision is made based on layers of voting schemes (Details will be illustrated in Subsection 3.2.)

3.1 Weighting Scheme

First of all, we partition each image into a number of equally sized non-overlapping regions in a consistent manner. In order to better illustrate this, we make the following assumptions and formalize a set of corresponding denotations.

Assuming we partition the image into $l \times m$ regions and name the region R(p,q) referring to the observation in row p and column q where $1 \leq p \leq l$ and $1 \leq q \leq m$. The regional scheme is shown in Figure 3.1.

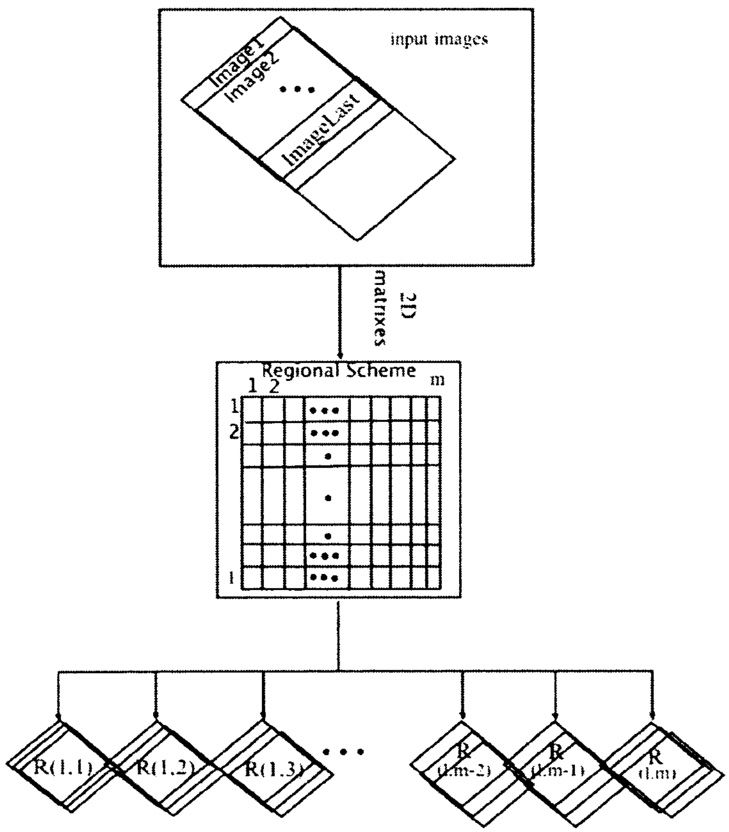


Figure 3.1: Regional Scheme

Assuming that, in our gallery, there are N subjects $S_1, S_2, \cdots, S_N$. Each subject, S_i, has K images $G_{i1}, G_{i2}, \cdots, G_{iK}$. There is a set of holistic algorithms $H =$

$\{h_1, h_2, \cdots, h_t\}$. For each holistic algorithm $h \in H$, on each region, we use a generated "leaving one out" strategy to test the effectiveness of the holistic algorithm on that region and take it as the weighting value of that region for algorithm h. For each j, $1 \le j \le K$, we select $G_{1j}, G_{2j}, \cdots, G_{Nj}$ as the testing set and take the remaining images in the gallery as the training set. By doing so, for each j, we find correctly recognized images for each region by that algorithm: $right(r_{(h,j)})$. The weighting value on region r for holistic algorithm h $w_{(h,r)}$ is calculated in an accumulated manner. It is formally defined in equation 3.1:

$$w_{(h,r)} = \frac{\sum_j {}_1^K right(r_{(h,j)})}{K \times N} \tag{3.1}$$

The weighting evaluation procedure on a region is shown in Figure 3.2. Thus, $w_{(h,r)}$ stand for the average recognition accuracy on region r by holistic algorithm h and $0 \le w_{(h,r)} \le 1$. Besides regional scheme, there are three key segments: "leaving one out" strategy, lower dimensional subspace and regional weights generator. Lower dimensional subspace includes a series of dimension reduction techniques, like holistic algorithms: PCA, Fisherface, SRDA, S-LDA and S-LPP in this case. The regional weights generator compares the subspace regional feature vectors by Euclidean distance and selects the closest one as the classification sticking to the nearest neighbour classifier. After calculating the regional weighting of holistic algorithm h according to equation 3.1, it implements one of the equations among equations 3.2, 3.3 and 3.4. During this stage, each region is an independent classifier.

The "leaving one out" strategy is shown in Figure 3.3. For each region, we use all the gallery images for training. The "leaving one out" strategy is used to divide the regional gallery images into subTrain and subTest sets. It includes K iterations of splitting[5]. To describe the "leaving one out" strategy, assume a k training dataset[6] with n different people in total. We denote the gallery images as:

$$\begin{bmatrix} the & first & image & of & subject1, & subject2, & \cdots & subjectn \\ the & second & image & of & subject1, & subject2, & \cdots & subjectn \\ & . & . & . & & & & \\ the & kth & image & of & subject1, & subject2, & \cdots & subjectn \end{bmatrix}$$

[5]splitting here refers to the division of gallery images into subTrain and subTest sets, not the split mentioned in the Chapter 4: Experiments, which refers to a component in the database.

[6]k refers to the number of images per subject for training (gallery).

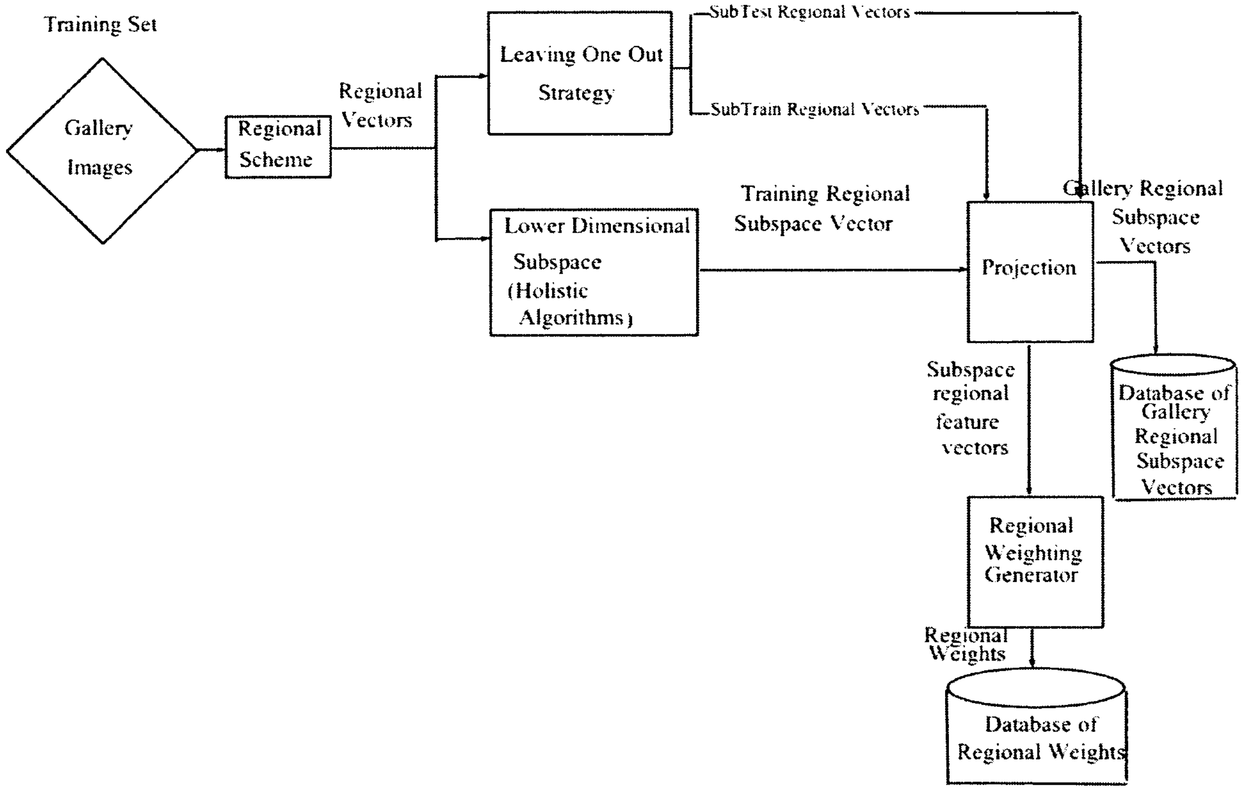


Figure 3.2: Weighting Evaluation

Each time, each row is "left in" and selected as the probe, all the rest are "left out" to become training images. Then during all rounds of splitting, we project the regional subTrain and subTest images into a subspace and find the recognition accuracies. Figure 3.3 shows splitting during one iteration. From the picture, we can see that, each time, the images of the same color are selected as the test images, keeping the rest as the training images.

In this thesis, five holistic algorithms are embedded in the framework and thus $t = 5$. h_1, h_2, h_3, h_4 and h_5 in turn correspond to algorithms: PCA, Fisherface, SRDA, S-LDA and S-LPP. With their weighting distributions over all regions of a face image, three different schemes are adopted in this thesis for the final weight to be used during the test stage after accepting probe images. We call them: "One Applies One", "One Applies All' and "Joint Weight" respectively. The following equations show the difference among the above three weighting schemes. In all cases, $w_{F(h,r)}$ stands for the final weight which is going to be assigned to the region r for holistic

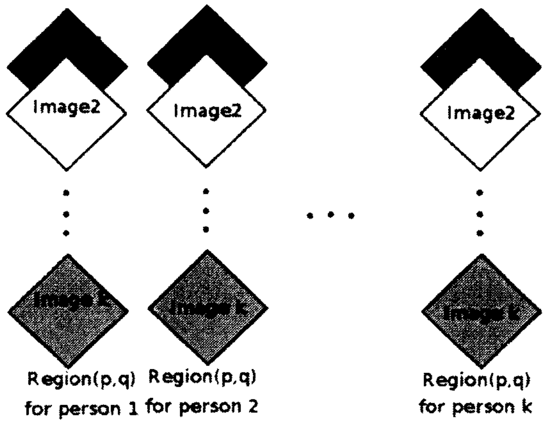


Figure 3.3: Splitting Iteration

algorithm h.

3.1.1 One Applies One

$$w_{F(h,r)} = w_{(h,r)} \tag{3.2}$$

3.1.2 One Applies All

$$w_{F(h,r)} = w_{(h_5,r)} \tag{3.3}$$

Here, during the weighting evaluation on the training set, only the effectiveness of S-LPP algorithm is tested. During the test stage having all probe images included, all algorithms (including S-LPP) use the weighting evaluated by S-LPP: $w_{(h_5,r)}$.

3.1.3 Joint Weight

$$w_{F(h,r)} = \sum_{al} {}_1^5 w_{(h_{al},r)} \tag{3.4}$$

where $w_{(h_{al},r)}$ refer to the weighting assigned to region r evaluated by algorithm h_{al} $\in$H.

3.2 Classifier Set Up and Recognition

This section involves two layers of voting/scoring models. The first layer is within the probe image. As we have a database of gallery regional vectors and a database of regional weights, for each holistic algorithm $h \in H$, we implement a regional weighting based classification of any probe image during the test stage. Euclidean distance is used during the matching procedure for comparison purpose and the smallest distance value identifies the subject ID of the probe image. Figure 3.4 describes this process.

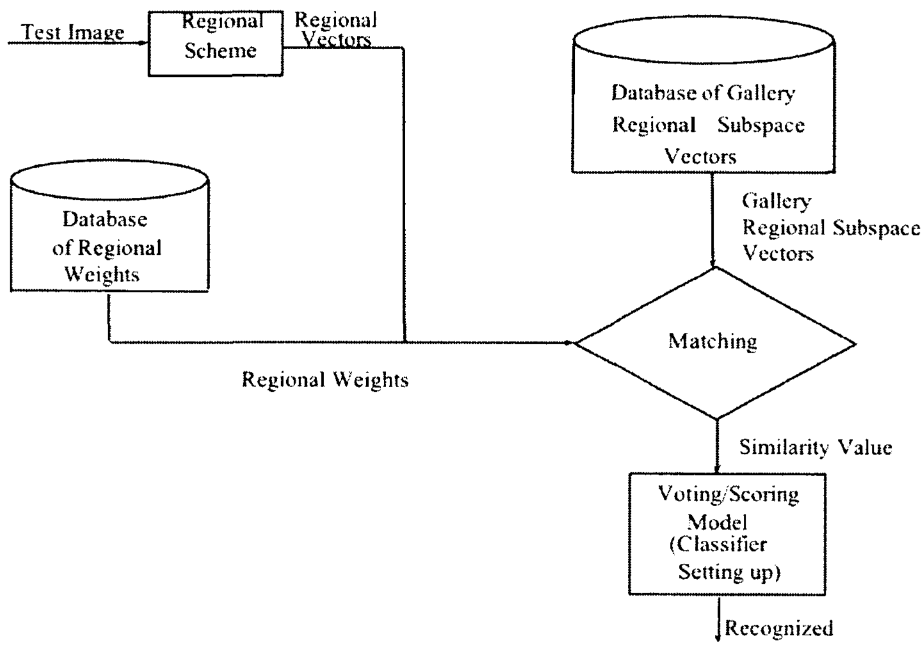


Figure 3.4: First layer Voting

The weighted scoring model is shown in Figure 3.5. To avoid non-meaningful comparisons, shifting is implemented. For even after pre-processing, pupils, mouth, chin, forehead etc. may still locate in different regions with respect to different face images. Thus, we perturb each gallery region up to two pixels in four directions (north, south, west and east) to compensate for misalignment issues. All regional gallery images in all (25 in total) nearby positions are compared with the regional probe image using a nearest neighbour classifier and the results are stored. By selecting the closest one as the identity of the holistic algorithm, h, on that region, we get the classification on region r: $h^r(p^r)$. Instead of one vote a region cast for an identity $i \in I$, $w_{F(h,r)}$ is used as the 'number' of votes identities get from each region. What the voting machine does is sum up the number of votes each identity gets. The one that gets the "biggest number" of votes is taken as the subject ID for probe image.

p, by that algorithm (h).

At this point, for each region r, for each holistic algorithm h, by using *all* gallery images, we obtain a classifier $h^r(p^r)$, where p is a probe image and $p \in P$. The classifier set up is shown in Table 3.1. Table 3.2 returns the corresponding classification for each classifier: $h^r(p^r)$. The identity that wins the biggest 'number' of votes is the final classification of holistic algorithm, h, on that probe image, p. By now, the first layer of voting for classification is set up.

Table 3.1: Classifier Set Up

$h^r p^r$ (1, 1)	$h^r p^r$ (1, 2)	$h^r p^r$ $\cdots$	h$^r p^r$ (1,m)
$h^r p^r$ (2, 1)	$h^r p^r$ (2, 2)	$h^r p^r$ $\cdots$	$h^r p^r$ (2,m)
$\vdots$	$\vdots$	$\cdots$	$\vdots$
$h^r p^r$ $(l, 1)$	$h^r p^r$ $(l, 2)$	$h^r p^r$ $\cdots$	$h^r p^r$ (l, m)

Table 3.2: Classification on each region by one holistic algorithm for the probe image

i_5	i_6	$\cdots$	i$_2$
i_3	i_1	$\cdots$	i$_2$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
i_2	i_n	$\cdots$	i$_{id}$

n in Table 3.2 is the total number of different identities to be classified and $1 \leq id \leq n$. Thus, for each holistic algorithm $h \in H$, we get its classification for each region.

The second layer of voting is among the different holistic algorithms. Each algorithm casts one vote to its identity and the final decision is made by a "winner takes all" strategy. That is, the identity voted by the majority of holistic algorithms is taken as the final result. After this round of voting, the final decision is made.

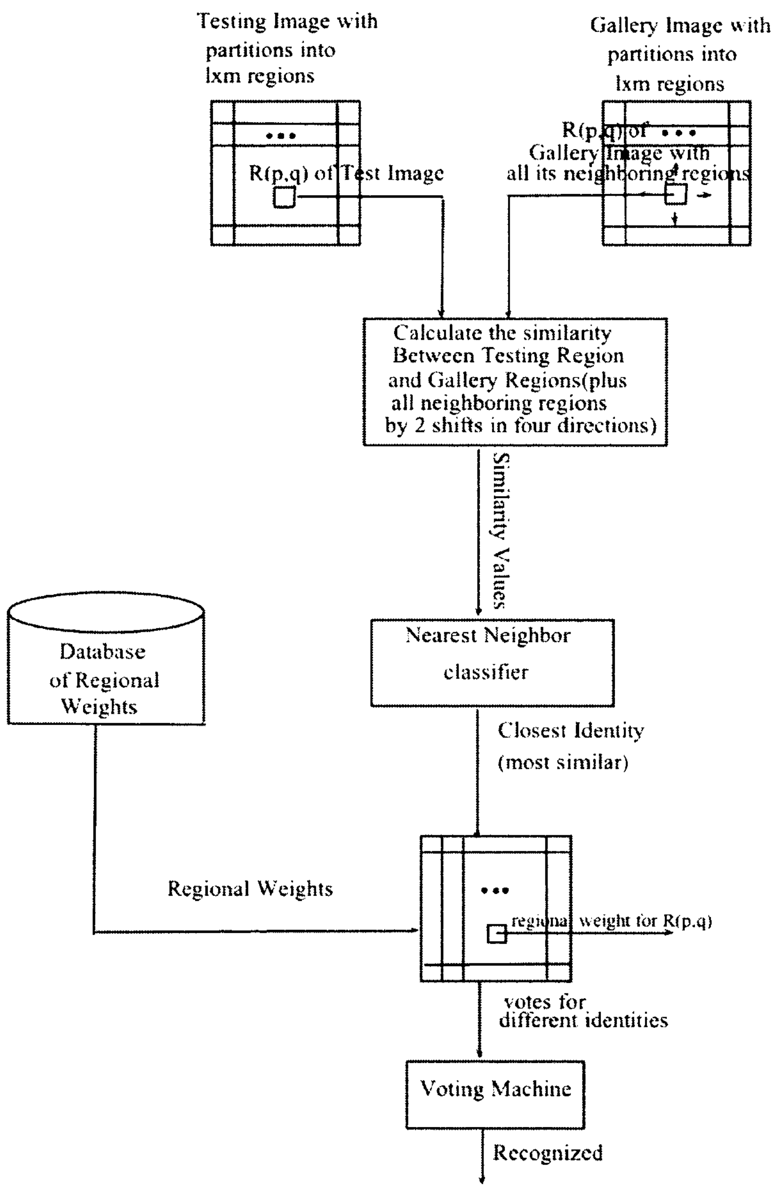


Figure 3.5: Voting Model

Chapter 4

Experiments

In order to validate the WREC approach, three benchmark databases were used. They decrease the technical difficulties with face recognition by:

- control of the environment such as background, lighting, camera angle and so on;
- control of the subject's pose;
- getting cooperation from subjects: having diverse images variant in expressions and accessories, wearing or not wearing glasses for instance.

The Yale database from Yale University contains 11 grayscale GIF images variant in facial expression and configuration - centre-light, with glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised, and wink[7] - of each of 15 individuals in total. Figure 4.1 is a glimpse of the Yale database[8].

Figure 4.1: Yale Face Database

The ORL database comes from Cambridge AT & T Laboratory, formerly Olivetti Research Laboratory. It contains ten different images of each of 40 distinct subjects. The images were taken at different times, varying the lighting, with different facial expressions (open / closed eyes, smiling / not smiling) and with different facial details (glasses / without glasses)[9]. Figure 4.2 comes from the ORL Database[10].

[7]See footnote 4
[8]See footnote 4
[9]See footnote 1
[10]See footnote 1

Figure 4.2: ORL Face Database

The Carnegie Melon University Pose, Illumination, and Expression database (CMU PIE) from Carnegie Melon University contains 41,368 images of 68 people. For each person, there are 13 different poses, 43 different illumination conditions and four different facial expressions[11]. Figure 4.3 is a glimpse of the PIE database[12].

Following the custom of researchers in this field, the faces for all three databases were simply manually aligned by pupils and cropped to 64×64 pixels with 256 gray levels per pixel. Here, I want to clarify that the alignment with respect to pupils as opposed to other features of the face, such as the nose or lip corners, is for the "state of art comparison". All the experiments are carried out exactly on the same data. As a matter of fact, though eye positions and inter-ocular distance are quite commonly used [P. Wang and Wayman, 2005], it does not mean that all pre-processing approaches have to wrap around the pupils. For instance, the Texas 3D Face Recognition Database (Texas 3DFRD)[13] has all faces normalized to the frontal position and the tip of the nose positioned at the center of the image[14]. It is difficult to standardize the face images with exactly the same techniques, thus it is hard to reproduce the face recognition algorithms of others to achieve the same performance. Therefore, it is fair to use published standardized data so that the comparison among the face recognition approaches is valid. To exclude any possible bias, including the pupil locating, rotating, scaling and cropping approach used for "standardization" during the Pre-Processing stage above mentioned, the UIUC versions of the ORL, Yale and PIE face databases are used. UIUC's version of the PIE database uses the near frontal poses (C05, C07, C09, C27, C29) which leaves us 11,554 face images. Each subject has 170 images, except subject 38 who has only 164 images. These

[11]See footnote 3

[12]See footnote 3

[13]`http://live.ece.utexas.edu/research/texas3dfr/`

[14]The description of the Texas 3DFRD could be found here: `http://www.face-rec.org/databases/`

pre-processed. scaled and rotated images were provided by Deng Cai[15].

Figure 4.3: PIE Face Database

Two baseline approaches (PCA and Fisherface) and three newly developed approaches (SRDA, S-LPP and S-LDA) are compared. For the Eigenface approach, the eigenvector corresponding to the largest eigenvalue was removed accounting for noise caused by illumination. $\alpha = 0.01$ was selected for regularization in SRDA and S-LPP approaches. For SLPP, cosine similarity was used to calculate the distances in the adjacency matrix. All gallery images were perturbed up to two pixels in each direction to make up for misalignment. All vectors(representation of images) were normalized to the length of 1. Probe images were linearly reduced and then classified according to nearest neighbour classification.

In the first set of experiments, images were divided d ($d = 11, 16$) times horizontally and vertically because these numbers yielded the best results with regional voting on the respective databases. And in this set of experiments. only the first weighting scheme was implemented. More specifically. in this set of experiments, WREC was using the "One Applies One" weighting scheme. The results for 2. 5 and 8 training datasets[16] on ORL and Yale databases are given in Tables 4.1 and 4.2 respectively. The results on the PIE Database for 5. 30. 80 and 130 training datasets are given in Table 4.3. In addition. a simple version of WREC named REC was also carried out in comparison. REC. as seen from the name. is a framework of WREC without weighting.

[15]All data and holistic algorithms were taken from `http://www.zjucadcg.cn/dengcai/Data/FaceData.html`.

[16]2. 5 and 8 refer to the number of images per subject for training (gallery). The rest data sets mentioned later in this thesis follow the same custom.

Table 4.1: WREC and REC in comparison with various face recognition algorithms on the Yale Database of 2, 5 and 8 training datasets with 11×11 and 16×16 divisions

	2 Train	5 Train	8 Train
Alg.	Error Rate±Std	Error Rate±Std	Error Rate±Std
PCA	44.40% ± 5.15%	33.84% ± 3.38%	30.93% ± 5.67%
Fisheface	43.02% ± 4.67%	10.73% ± 3.16%	7.24% ± 3.41%
SRDA	30.71% ± 4.69%	11.38% ± 2.96%	6.80% ± 4.06%
S-LDA	29.90% ± 5.09%	13.58% ± 3.11%	8.98% ± 4.17%
S-LPP	32.50% ± 4.40%	12.93% ± 3.60%	8.22% ± 4.38%
REC(11)	13.99% ± 2.54%	5.33% ± 2.45%	2.90% ± 2.24%
WREC(11)	11.16% ± 2.37%	4.26% ± 2.10%	2.18% ± 2.02%
REC(16)	13.57% ± 2.79%	5.29% ± 2.44%	2.51% ± 2.01%
WREC(16)	10.26% ± 2.76%	4.10% ± 1.95%	1.91% ± 2.01%

In the second set of experiments, conducted on the Yale and ORL datasets, WREC were carried out with different number of regions, from 7×7 to 20×20. That is, each image was first divided 7 times vertically, and 7 times horizontally, and then 8 times each direction and so on up to 20 divisions vertically and 20 horizontally. This time, all three weighting schemas were implemented in this set of experiments. The results for the Yale database on 2, 5 and 8 training datasets are given in Figures 4.4, 4.5 and 4.6 respectively. And the results for the ORL database on 2, 5 and 8 training dataset are given in Figures 4.7, 4.8 and 4.9 respectively.

On the PIE database, due to time constraints as well as the possible distribution of the best results, the experiments were carried out on divisions from 7 up to 16 with only one weighting scheme: One Applies One. As the recognition accuracy for the 5 Train dataset differs a lot from the rest of the datasets, we put it aside in a separate figure to have a better representation of the result. The result is shown in Figure 4.10 and 4.11.

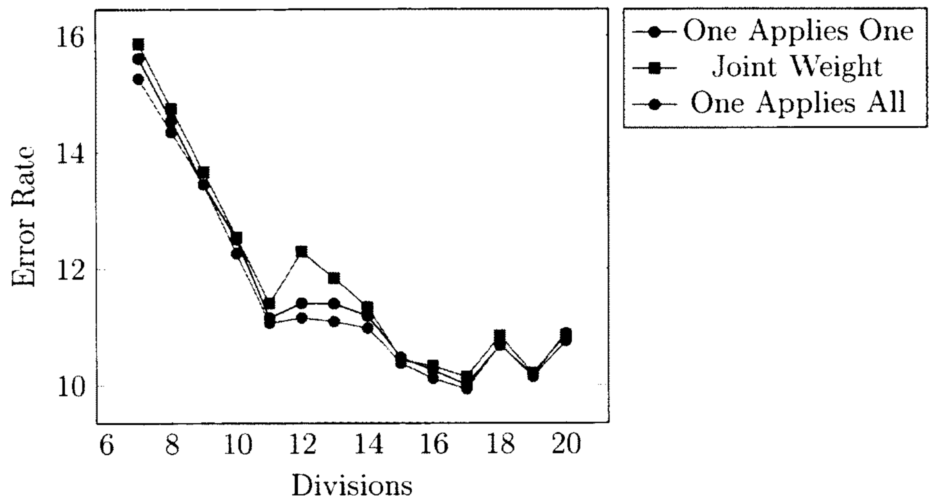


Figure 4.4: WREC on Yale database (2 training dataset) with divisions from 7 up to 20

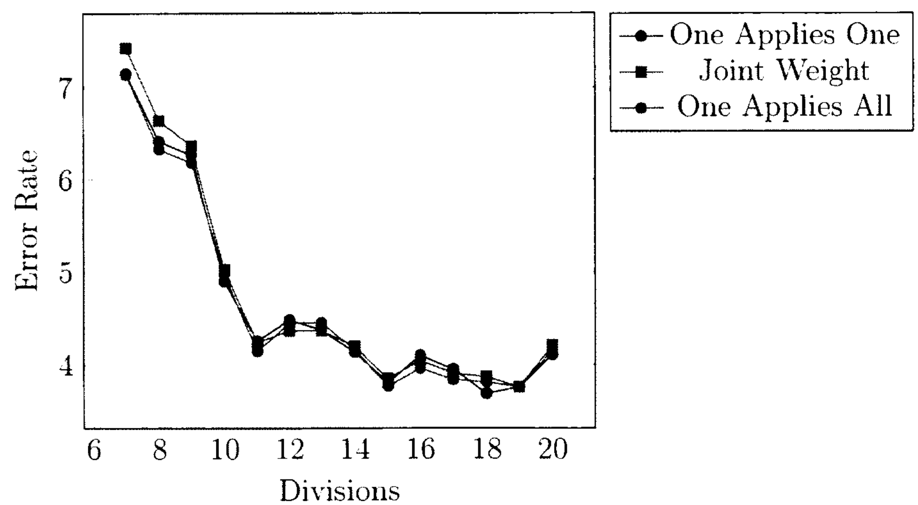


Figure 4.5: WREC on Yale database (5 training dataset) with divisions from 7 up to 20

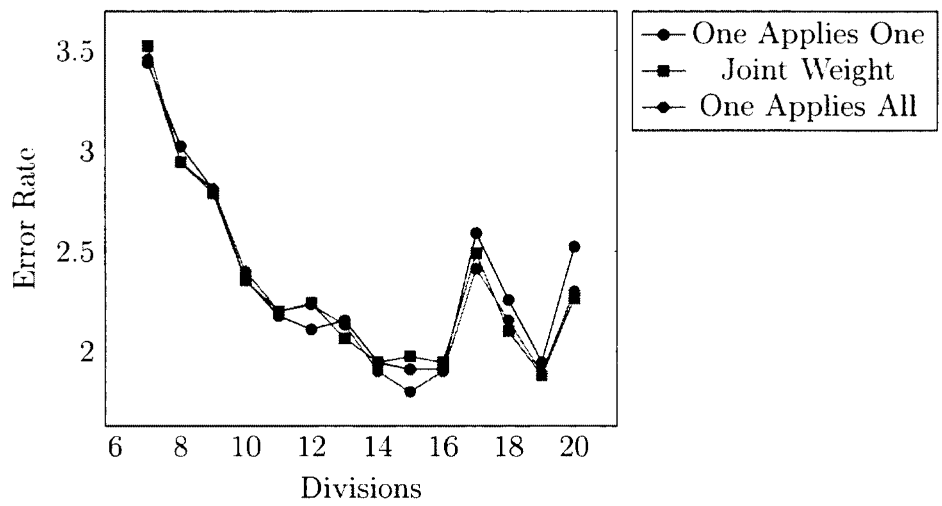


Figure 4.6: WREC on Yale database (8 training dataset) with divisions from 7 up to 20

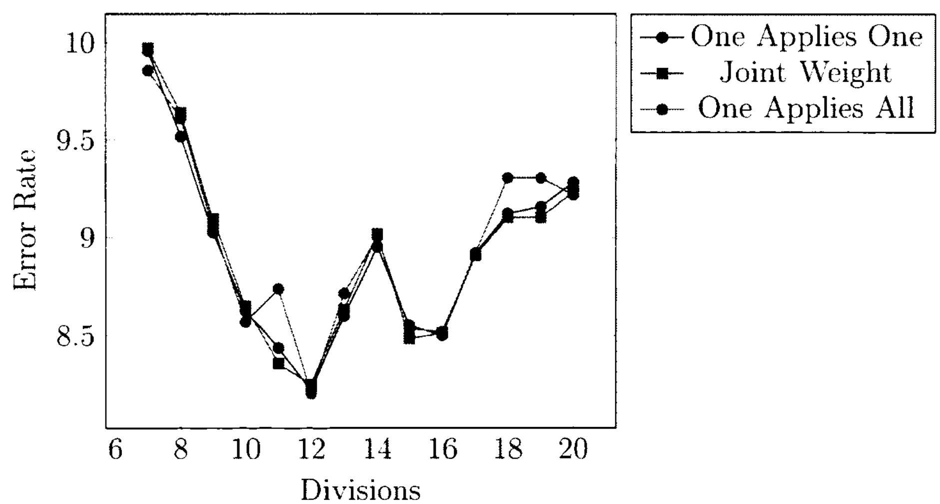


Figure 4.7: WREC on ORL database (2 training dataset) with divisions from 7 up to 20

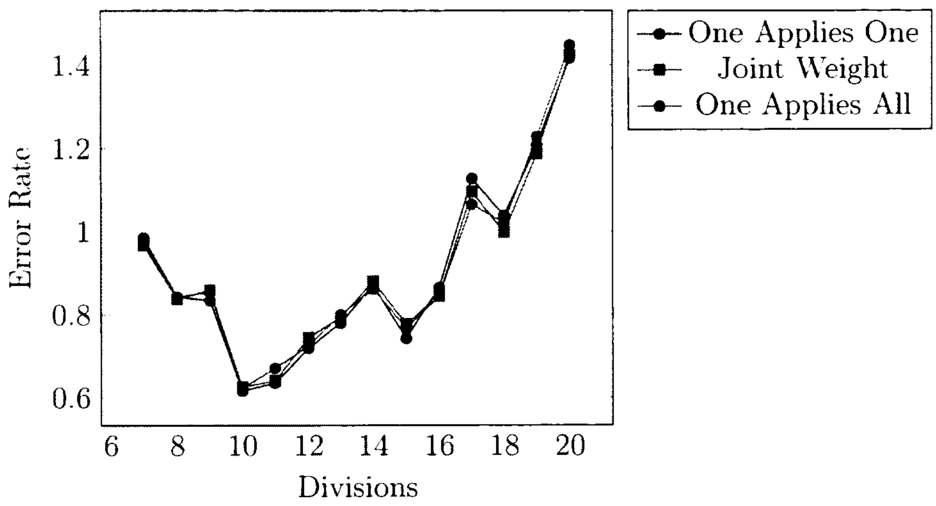


Figure 4.8: WREC on ORL database (5 training dataset) with divisions from 7 up to 20

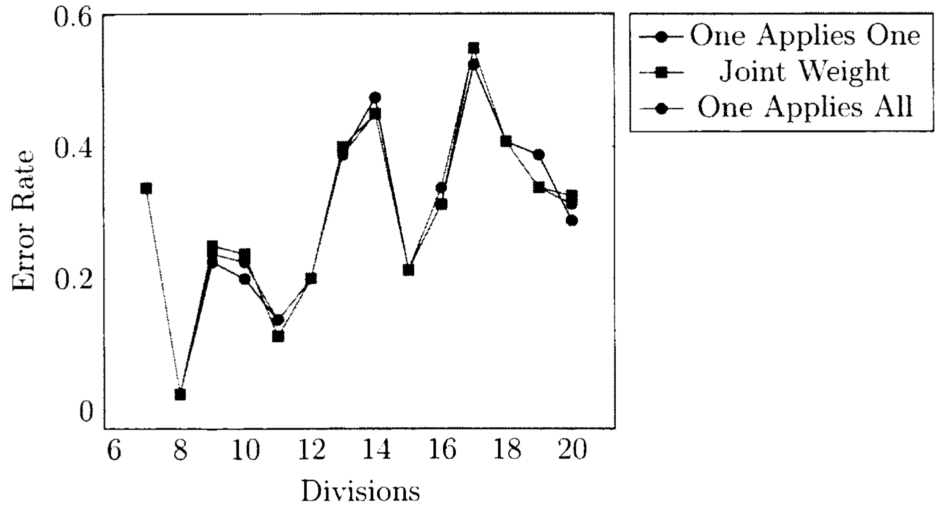


Figure 4.9: WREC on ORL database (8 training dataset) with divisions from 7 up to 20

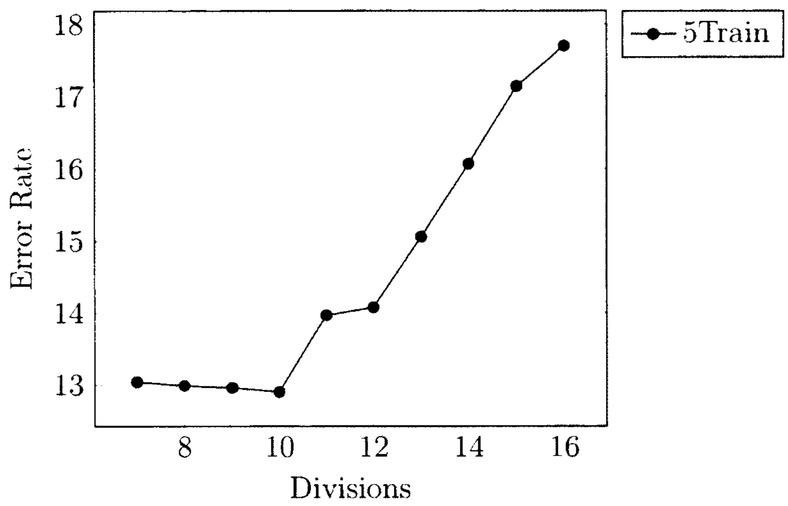


Figure 4.10: WREC on PIE database (5 training dataset) with divisions from 7 up to 16

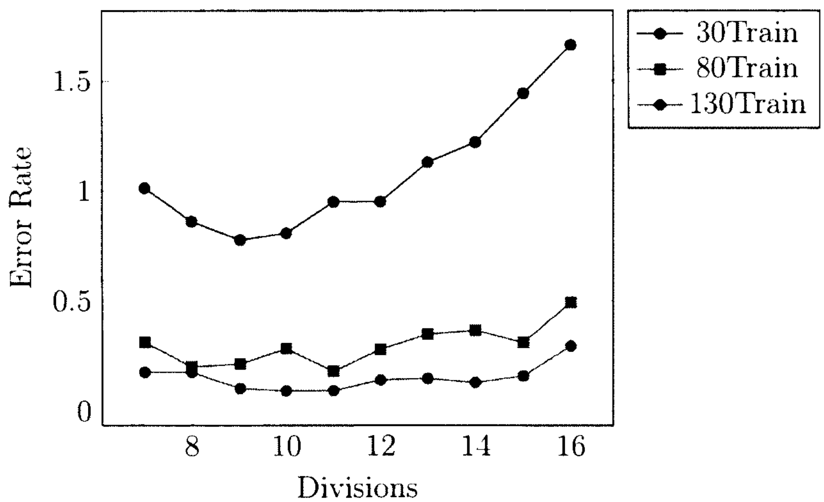


Figure 4.11: WREC on PIE database (30, 80 and 130 training datasets) with divisions from 7 up to 16

Table 4.2: WREC and REC in comparison with various face recognition algorithms on the ORL Database of 2, 5 and 8 training datasets with 11×11 and 16×16 divisions

	2 Train	5 Train	8 Train
Alg.	Error Rate±Std	Error Rate±Std	Error Rate±Std
PCA	$29.29\% \pm 3.15\%$	$11.48\% \pm 2.26\%$	$6.05\% \pm 2.27\%$
Fisheface	$22.28\% \pm 2.82\%$	$3.45\% \pm 1.30\%$	$1.65\% \pm 1.18\%$
SRDA	$18.19\% \pm 2.81\%$	$3.44\% \pm 1.19\%$	$1.80\% \pm 1.50\%$
S-LDA	$16.45\% \pm 2.92\%$	$2.47\% \pm 1.08\%$	$0.83\% \pm 1.14\%$
S-LPP	$17.23\% \pm 3.04\%$	$2.62\% \pm 1.20\%$	$0.93\% \pm 0.99\%$
REC(11)	$8.84\% \pm 2.05\%$	$0.81\% \pm 0.85\%$	$0.16\% \pm 0.39\%$
WREC(11)	$8.44\% \pm 1.90\%$	$0.64\% \pm 0.78\%$	$0.14\% \pm 0.38\%$
REC(16)	$8.97\% \pm 2.14\%$	$0.96\% \pm 0.75\%$	$0.43\% \pm 0.58\%$
WREC(16)	$8.50\% \pm 2.02\%$	$0.87\% \pm 0.76\%$	$0.31\% \pm 0.52\%$

Table 4.3: WREC and REC in comparison with various face recognition algorithms on the PIE Database of 2, 5 and 8 training datasets with 11×11 and 16×16 divisions

	5 Train	30 Train	80 Train	130 Train
Alg.	Error Rate	Error Rate	Error Rate	Error Rate
PCA	70.23%	26.53%	7.18%	2.43%
Fisheface	30.28%	12.04%	8.16%	5.71%
SRDA	28.32%	5.01%	3.12%	2.65%
S-LDA	25.82%	3.50%	1.83%	1.58%
S-LPP	27.71%	4.79%	2.58%	1.66%
REC(11)	15.63%	1.01%	0.24%	0.13%
WREC(11)	13.97%	0.95%	0.18%	0.09%
REC(16)	20.19%	1.79%	0.57%	0.31%
WREC(16)	17.71%	1.62%	0.49%	0.29%

Chapter 5

Analysis

The experiments demonstrate WREC's significant performance advantages compared to several other leading approaches. The results shown in the tables achieve overwhelmingly better recognition accuracy than all the other algorithms in all cases over all datasets. In a lot of cases, the error recognition rate drops more than half. Figures 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 and 4.11, all show the same pattern as Regional Voting [Chen and Tokuda, 2005] that accuracy goes up as the number of regions increases. After a certain point, the accuracy begins to drop for the regions become too small to distinguish from national voting. The above observations from the figures match precisely the theory of "Electoral College and Direct Popular vote".

Besides, Figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9 and 5.10 show the transition point (the division where the recognition performance peaks) in WREC system appear earlier than Regional Voting. In this set of figures, WREC is represented by using the "One Applies One" weighting scheme for face recognition and accuracy does not differ much from the variant weighting versions for WREC. Even though, for each region, more algorithms are involved, we can achieve the best result within a shorter time. From the tables, we can see that the pure ensemble of multiple classifiers without weights also outperforms standard regional voting.

Regarding the acceptable error rate for face recognition, we have to keep in mind that it is not governed by a formal theory. Different from that in physics or mathematics, there is no absolute cutoff value (an error rate that can be applied to all other applications). Face recognition, although it has been an active research topic for more than 20 years, is still not mature enough to have a generally acceptable standard. The statistical success and failures are application-dependent. [R. Chellappa and Sirohey, 1995].

Figure 5.11 shows the weighting distribution for 16×16 divisions on the Yale dataset (2 Training Dataset) by different classifiers.

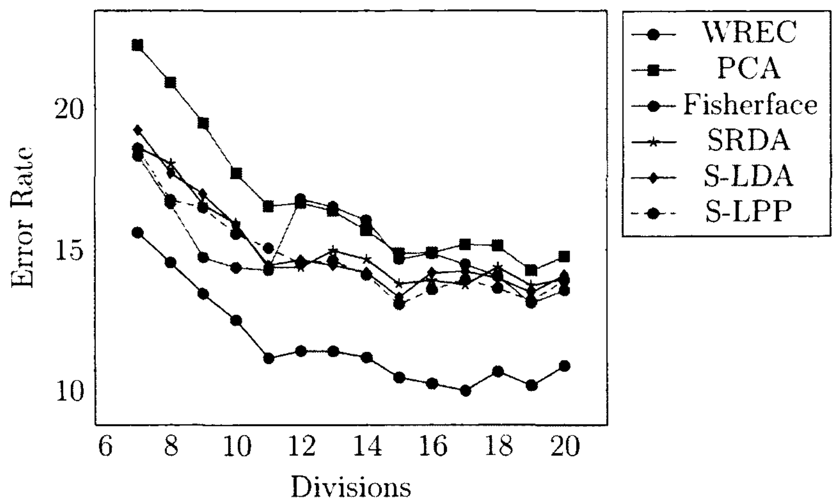


Figure 5.1: WREC compared to various individual holistic algorithms in different sized regions on Yale 2 training dataset

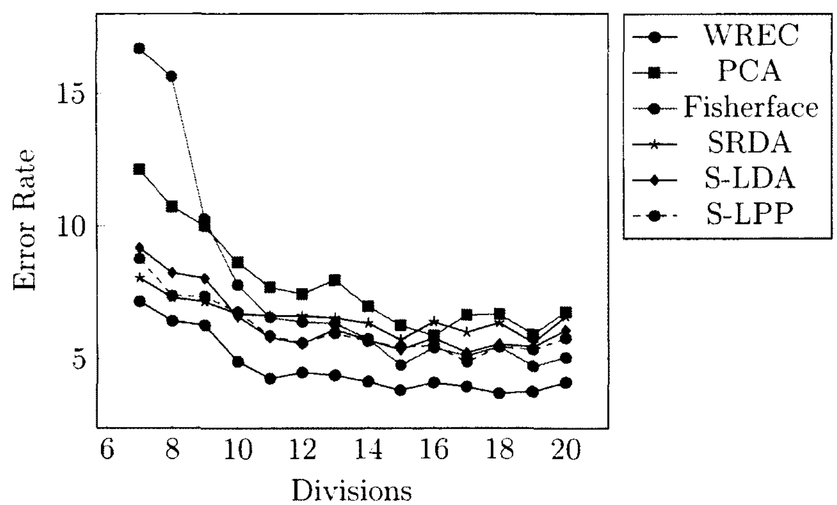


Figure 5.2: WREC compared to various individual holistic algorithms in different sized regions on Yale 5 training dataset

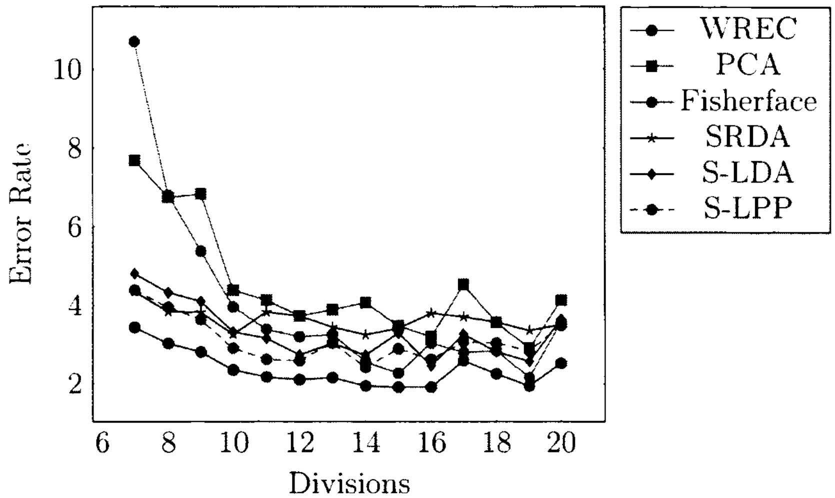


Figure 5.3: WREC compared to various individual holistic algorithms in different sized regions on Yale 8 training dataset

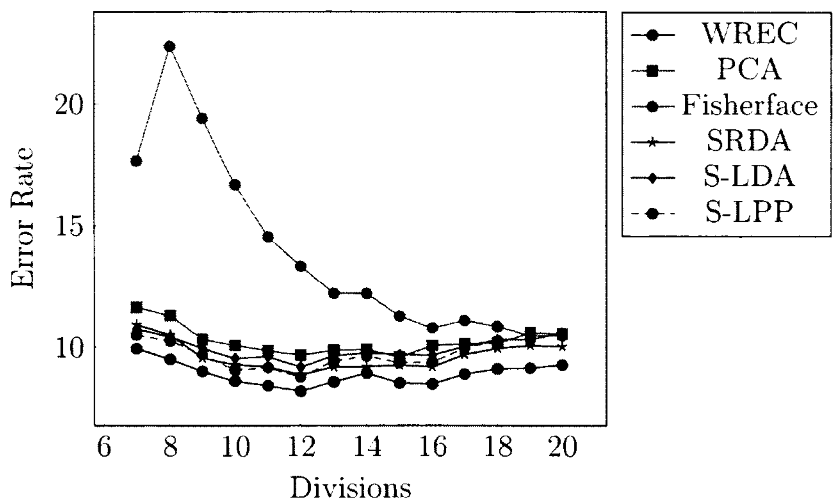


Figure 5.4: WREC compared to various individual holistic algorithms in different sized regions on ORL 2 training dataset

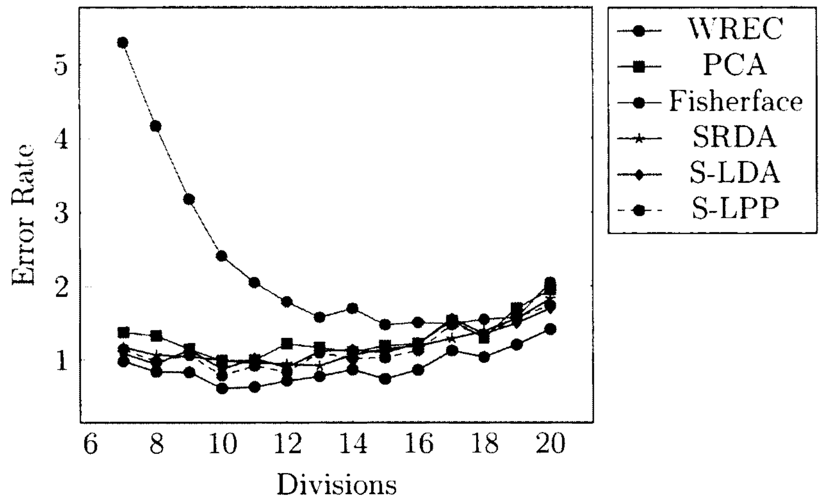


Figure 5.5: WREC compared to various individual holistic algorithms in different sized regions on ORL 5 training dataset

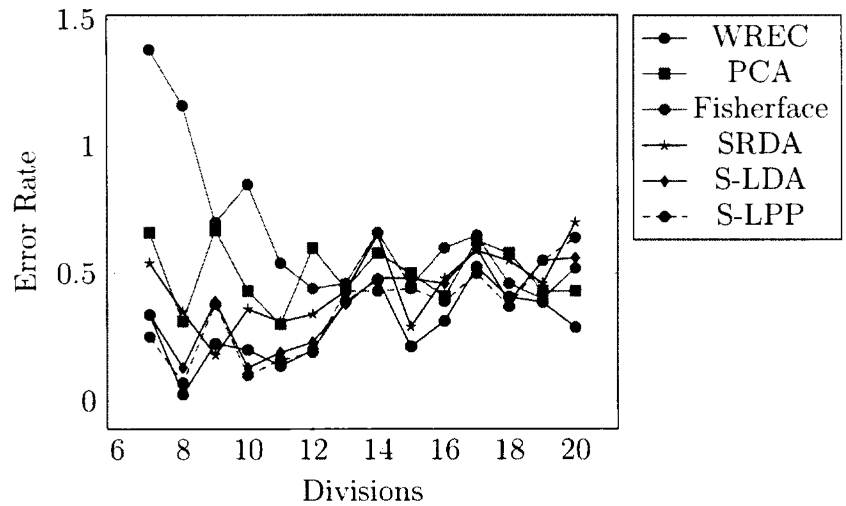


Figure 5.6: WREC compared to various individual holistic algorithms in different sized regions on ORL 8 training dataset

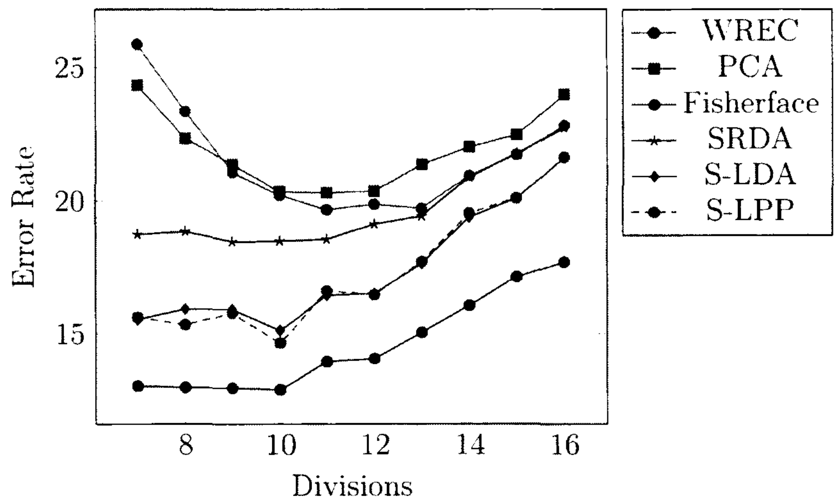


Figure 5.7: WREC compared to various individual holistic algorithms in different sized regions on PIE 5 training dataset

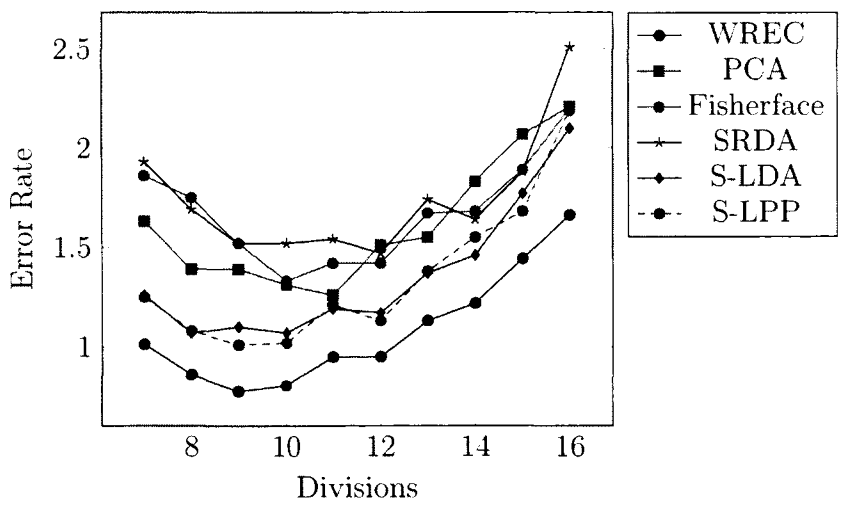


Figure 5.8: WREC compared to various individual holistic algorithms in different sized regions on PIE 30 training dataset

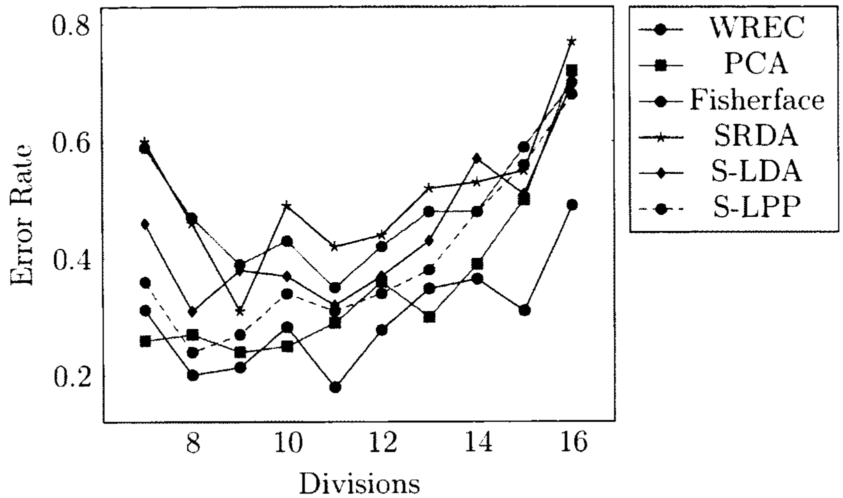


Figure 5.9: WREC compared to various individual holistic algorithms in different sized regions on PIE 80 training dataset

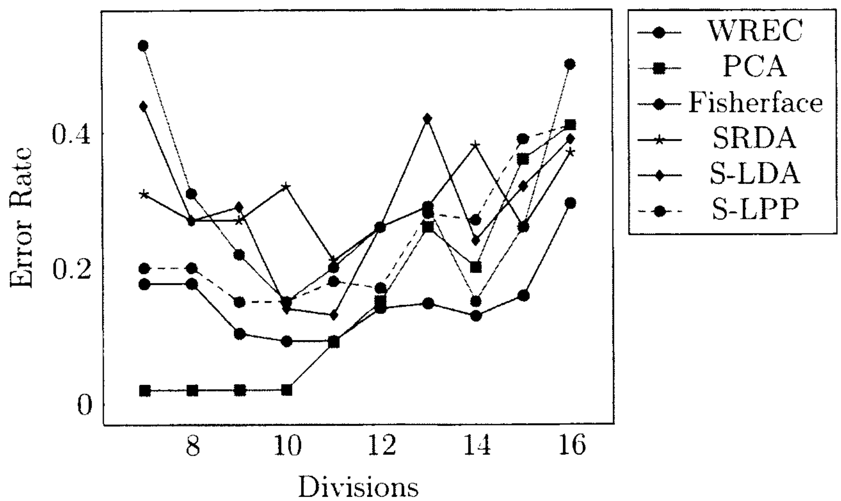


Figure 5.10: WREC Compared to various individual holistic algorithms in different sized regions on PIE 130 training Dataset

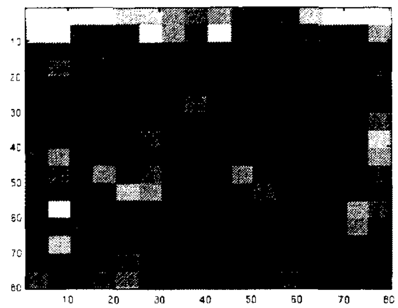
Figure 5.12 shows the weighting distribution for 16×16 divisions on the Yale dataset (8 training dataset) this time. The more faces involved in calculating the weighting, the closer it should be able to derive the face features.

The brighter the region is, the higher the weight associated with that region. The figures, either Figure 5.11 or Figure 5.12 which draw weighting distributions from a larger pool of samples, tell us that salient features like eyes, mouth and nose do not necessarily yield higher weights. This confirms the superiority of using an automatic weighting estimation technique. Different weighting schemes have been proposed [Xiaoyan Mu and Watta, 2005, Timo Ahonen and Pietikainen, 2004] and in this thesis, an automatic weighting system is adopted. Thus, unlike the weighting algorithm Timo Ahonen et al proposed [Timo Ahonen and Pietikainen, 2004], which produced a symmetric weight distribution on classifiers, it treats each region separately. Combining classifiers mentioned in Chapter 2, should work better than having symmetric weights attached to the regions.

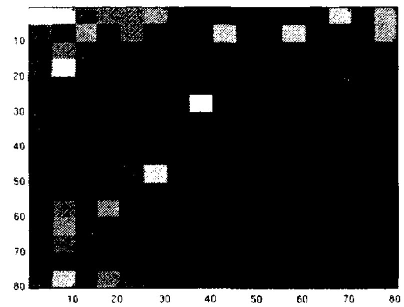
Finally, regarding time complexity, obviously, it is dependent on that of the embedded algorithms. Despite the fact of the disparity in complexities for different holistic algorithms, we denote the time cost for each holistic algorithm as τ. Then the time complexity of Regional Voting is $\iota\tau$ where ι is the number of regions. Examining each dimension-reduction algorithm, the time complexity of τ is made up of the matrix multiplication with time complexity of $O(\iota hw\kappa)$ where h and w are the height and width of the image and κ is the number of dimensions of the reduced subspace. After that, a classification based on the measurement of the distances between probe image and each gallery image (including all its neighboring images by shifting) is performed. During this stage, the time complexity is $O(\iota hw\rho n)$ where ρ is the total number of shifts and n is the total number of different subjects to be classified. Assume we have t holistic algorithms embedded in total. So the total complexity is $O(t\iota hw(\kappa + \rho n))$ in the test stage. Examining the time complexity of the training, it is much like the that of the test stage except multiplying by k, where k is the number of images per person in the gallery, since a "leaving one out strategy" is implemented. For each algorithm, after k iterations, $k\iota hw(\kappa + \rho n)$ operations (shifting is again applied in the training) are required.

Of the three different weighting schemes we have implemented in this thesis. "One

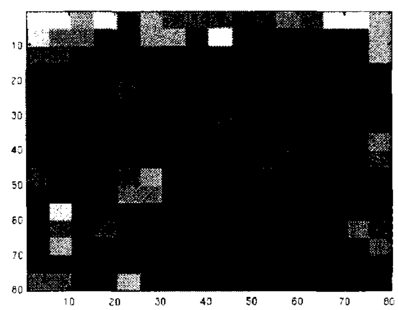
Applies One", "One Applies All" and "Joint Weight" - for both "One Applies One" and "Joint Weight" schemes, all holistic algorithms embedded are used. Thus the time complexity is $tk\iota hw(\kappa + \rho n)$. And lastly, the "One Applies All" weighting scheme shortens the training period roughly to 1/5 as long for only one holistic algorithm is used in the training stage instead of five. The complexity is down to $k\iota hw(\kappa+\rho n)$. To this scheme's credit is that face recognition accuracy still runs neck to neck compared with that of the other two schemes.



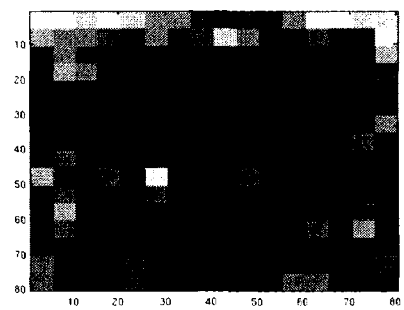
(a) Weighting Distribution by PCA



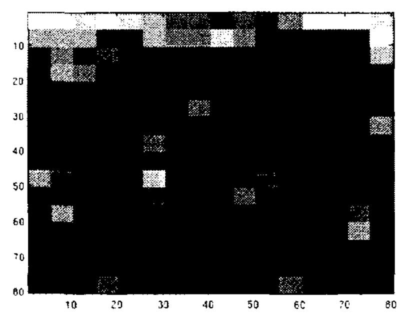
(b) Weighting Distribution by Fisherface



(c) Weighting Distribution by SRDA

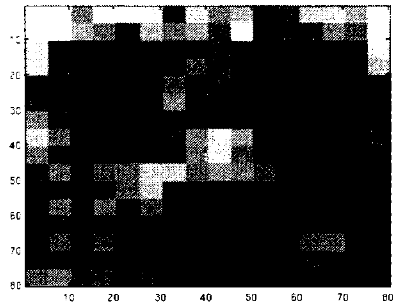


(d) Weighting Distribution by S-LDA

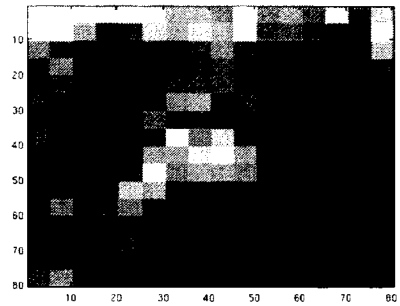


(e) Weighting Distribution by S-LPP

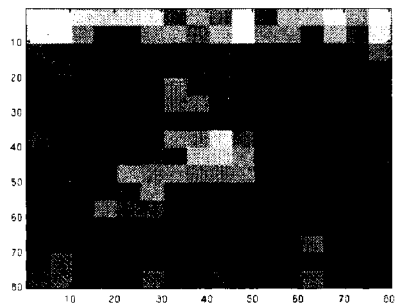
Figure 5.11: Weighting Distribution by Different Classifiers on 16×16 division of Yale Data Set(2 training dataset)



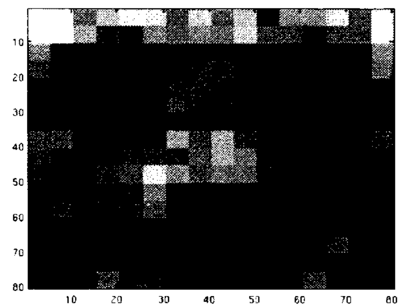
(a) Weighting Distribution by PCA



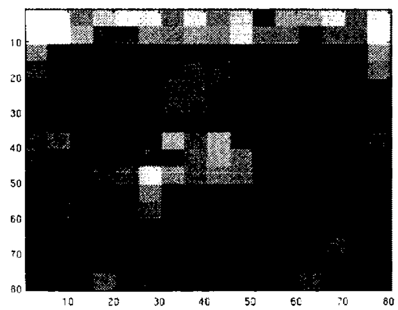
(b) Weighting Distribution by Fisherface



(c) Weighting Distribution by SRDA



(d) Weighting Distribution by S-LDA



(e) Weighting Distribution by S-LPP

Figure 5.12: Weighting Distribution by Different Classifiers on 16×16 division of Yale Data Set(8 training dataset)

Chapter 6

Summary

Face recognition is an emerging technology. Though it is by no means perfect, there is no denying its tremendous potential. Compared to human facial surveillance, automatic systems have a longer attention span, can be exposed to larger amounts of information and work in the same manner reliably in all cases. Holistic approaches, based on the concept of reducing the high dimensionality of the raw face image are regarded as the most popular solution for the face recognition task.

Regional voting is a new approach for the face recognition problem which divides the image into equally sized non-overlapping regions and treats each region as an independent classifier. Within each region, holistic algorithms are implemented and the classification result is recorded. After all, a vote over all regions is carried out and the classification with majority votes is selected as the final result.

WREC builds on the success of Regional Voting in the following way:

1. The Electoral College framework is adopted into the proposed system successfully.

2. An automatic weighting calculation method independent of human knowledge is implemented. This enables a more robust system accommodating face images having a bigger range of variance; for example, face images having non-symmetric face features violated by postures, face expressions or occlusions.

3. Three different schemes for weighting are proposed.

4. In the newly proposed system, embedded algorithms are not independent anymore. Two layers of voting models are included in the system.

5. Extensive experiments are carried out on benchmark face databases. The proposed system shows superiority over several other leading algorithms as well as the already best in class results of regional voting. The same promising results

are also derived on experiments of datasets with small number of images per person in the gallery.

WREC, as a solution to the face recognition problem, is validated by extensive experiments, outperforming any individual holistic approaches embedded. Even a REC (a regional based ensemble of multiple classifiers) system shows promising results on datasets with a smaller number of gallery images.

Finally, there are some points regarding the design of this experiment. This may suggest avenues for future work. Firstly, during the splitting, after each round, we get two sets of images: SubTest and SubTrain. Figure 6.1 shows the division on part of the ORL 8 training dataset picking up the first image of a person as the SubTest images. This splitting is easy to implement, while it does not include all objects that can be used for matching. The first image of the person himself is excluded as there is no distance at all between one and himself, while the other people's first image are also excluded by this algorithm. Both images come from the ORL Database[17].

(a) Sub-Test Images

(b) SubTraining Images

Figure 6.1: Example of regional divisions on gallery images

So, in the future, we could include more information provided by the gallery images during weighting evaluation to test the effectiveness of an algorithm on each region. Also, we can embed a wider range of algorithms, like the ones under local branch, for example. And lastly, the system could be expanded for face verification.

[17]See footnote 1.

Bibliography

Nasir U. Ahmed and K. Ramamohan Rao. *Orthogonal Transforms for Digital Signal Processing*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1975.

K.M. Ali and M.J. Pazzani. On the Link Between Error Correlation and Error Reduction in Decision Tree Ensembles. Technical report, 95-38, ICS-UCI, 1995.

G. H. Golub B. L. Buzbee and C. W. Nielson. On Direct methods for solving poisson's equations. *SIAM Journal on Numerical Analysis*, 7(4):627–656, 1970.

S.C. Bagui and N.R. Pal. A Multistage Generalization of the Rank Nearest Neighbor Classification Rule. *Pattern Recognition Letters*, 16:601–614, 1995.

Hespanha. J Belhumeur. P and Kriegman. D. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

M. Bichsel and A.P. Pentland. Human face recognition and the face image set's topology. *CVGIP: Image Understanding*, 59:254–261, 1994.

Deng Cai, Xiaofei He, and Jiawei Han. Semi-supervised Discriminant Analysis. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–7. IEEE, 2007a.

Deng Cai, Xiaofei He, and Jiawei Han. Spectral Regression for Efficient Regularized Subspace Learning. *IEEE 11th International Conference on Computer Vision (2007)*, L(05):1–8, 2007b.

Y. Hu J. Han Cai D., X. He and T. Huang. Learning a spatially smooth subspace for face recognition. *IEEE International Conference on Computer Vision Pattern Recognition*, 0:1–7, 2007.

He Xiaofei Cai Deng and Han Jiawei. SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):1–12, 2008.

Liang Chen and Naoyuki Tokuda. Robustness of regional matching scheme over global matching scheme. *Artificial Intelligence*, 144(1-2):213–232, 2003.

Liang Chen and Naoyuki Tokuda. A general stability analysis on regional and national voting schemes against noise–why is an electoral college more stable than a direct popular election? *Artificial Intelligence*, 163(1):47–66, 2005.

Liang Chen and Naoyuki Tokuda. A unified framework for improving the accuracy of all holistic face identification algorithms. *Artif. Intell. Rev.*, 33(1-2):107–122, 2010.

Ravi Das. An Application of Biometric Technology: Facial Recognition: http://www.technologyexecutivesclub.com/Articles/security/ artBiometricsFacial-Recognition.php, 2011.

R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

H. D. Ellis. *Processes underlying face recognition: Introduction in The Neurophysiology of Face perception and Facial Expression.* Hillsdale, NJ: Erlbaum, 1986.

Martha J. Farah. Is face recognition 'special' Evidence from neuropsychology. *Behavioural Brain Research*, 76:181–189, 1996.

Ronald Aylmer Fisher. The Use of Multiple Measures in Taxonomic Problems. *Annals of Eugenics*, 7:179–188, 1936.

Rafael C Gonzalez, Richard E Woods, and Barry R Masters. Digital Image Processing, Third Edition. *Journal of Biomedical Optics*, 14(2):029901, 2009.

H.-W. Chang H.-T. Chen and T.-L. Liu. Local discriminant embedding and its variants. In *Proc. CVPR*, 2005.

Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood Preserving Embedding. In *Tenth IEEE International Conference on Computer Vision ICCV05 Volume 1*, volume 2, pages 1208–1213. Ieee, 2005.

Xiaofei He and Partha Niyogi. Locality Preserving Projections. *Advances in Neural Information Processing Systems*, 16:153–160, 2003.

Hu Yuxiao Niyogi Partha He Xiaofei, Yan Shuicheng and Zhang Hongjiang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

Rebecca Heyer. *Biometrics Technology Review 2008.* Land Operations Division Defence Science and Technology Organisation, 2008.

R. Horn and C. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

Y Ijiri and M Sakuragi. Security Management for Mobile Devices by Face Recognition. In *7th International Conference on Mobile Data Management MDM06*, pages 49–49, 2006.

A. N. Venetsanopoulos Juwei Lu, K. N. Plataniotis and Stan Z. Li. Ensemble-based discriminant learning with boosting for face recognition. *IEEE Transactions on Neural Networks*, 17:178, 2006.

Takeo Kanade. Computer recognition of human faces. *Interdiscilplinary Sys- tems Research*, page 47, 1977.

Hatef Mohamad Duin Robert P. W. Kittler, Josef and Jiri Matas. On Combining Classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:226–239, 1998.

T. Kohonen. *Self-Organization and Associative Memory.* Springer, 1988.

J. M. Lee. *Introduction to Smooth Manifolds.* Springer-Verlag New York, 2002.

J. Buhmann J. Lange C. v. d. Malsburg M. Lades, J. Vorbruggen and R. Wurtz. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42:300–311, 1993.

Crookes K. McKone E. and Kanwisher N. *The cognitive and neural development of face recognition in humans.* Cambridge, MA, Bradford Books, 2009.

Christian A Meissner and John C Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology Public Policy and Law*, 7(1):3–35, 2001.

Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception*, 30:303–321, 2001.

John Morton and Mark H. Johnson. Conspec and Conlearn: A Two-Process Theory of Infant Face Recognition. *Psychological Review*, 98, No.2:164–181, 1991.

F. O'Sullivan. Discretized laplacian smoothing by fourier methods. *Journal of the American Statistical Association*, 86(415):634–643, 1991.

L. Ganesan P. Latha and S. Annadurai. Face Recognition using Neural Networks. *Signal Processing: An International Journal (SPIJ)*, 3:153–160, 2009.

Q. Ji P. Wang, M.B. Green and J. Wayman. Automatic eye detection and its validation. *In Proc. of 2005 IEEE Conf. on Computer Vision and Pattern Recognition - Workshops*, 3:164, 2005.

Christopher C Paige and Michael A Saunders. LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. *ACM Transactions on Mathematical Software*, 8(1):43–71, 1982.

Olivier Pascalis and Jocelyne Bachevalier. Face recognition in primates: a cross-species study. *Behavioural Processes*, 43(1):87–96, 1998.

P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation. *Netw.: Comput. Neural Syst.*, 7:477–500, 1996.

Alex Pentland and Tanzeem Choudhury. Face recognition for smart environments. *Computer*, 33(2):50–55, 2000.

Rauss P Phillips P J, Moon H and Rizvi S. The FERET evaluation methodology for face recognition algorithms. In *Proceedings of Computer Vision and Pattern Recognition 97*, 1997.

C.L. Wilson R. Chellappa and S. Sirohey. Human and Machine Recognition of Faces: A Survey. *Proceedings of the IEEE*, 83:705–740, 1995.

S.J. Raudys and A.K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13: 252–264, 1991.

D. L. Ruderman. The statistics of natural images. *Netw.: Comput. Neural Syst.*, 5: 598–605, 1994.

J. Shlens. A Tutorial on Principal Component Analysis. Technical report, Institute for Nonlinear Science, UCSD, 2005.

T. J. Stonham. *Practical face recognition and verication with wisard.* Aspects of face processing,Kluwer Academic Publishers, 1986.

R Veldhuis R Gehlen S Tao Q, Van Rooseler and Weber. Optimal Decision Fusion and Its Application on 3D Face Recognition: http://www.3dface.org/files/papers/veldhuis-CAST2007OptimalDecisionFusion.pdf, 2007.

Abdenour Hadid Timo Ahonen and Matti Pietikainen. Face recognition with local binary patterns. *Lecture Notes in Computer Science*, 3021/2004:469–481, 2004.

J.J. Hull T.K. Ho and S.N. Srihari. Decision Combination in Multiple Classifier Systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16:66–75, 1994.

D Tranel and AR Damasio. Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics. *Science*, 228:1453–1454, 1985.

Matthew Turk and Alex Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

T Valentine. Upside-down faces: a review of the effect of inversion upon face recognition. *British journal of psychology London England 1953*, 79 (Pt 4)(4):471–491, 1988.

Xiaogang Wang and Xiaoou Tang. Unified Subspace Analysis for Face Recognition. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2: 679686, 2003.

Mohamad Hassoun Xiaoyan Mu and Paul Watta. A Weighted Voting and Sequential Combination of Classifiers Scheme for Human Face Recognition,. In *International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, 2005.

Y. Adini Y. Moses and S. Ullman. Face recognition: The problem of compensating for change in illumination direction. In *In European Conf. On Computer Vision*, 1994.

S Yan, D Xu, B Zhang, and Hong-Jiang Zhang. Graph Embedding: A General Framework for Dimensionality Reduction. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05*. volume 2, pages 830–837. Ieee, 2005.

W Tan J Jia Z Zheng, F Yang and J Yang. Gabor feature-based face recognition using supervised locality preserving projection. *Signal Processing*, 87(10):24732483, 2007.

W Zhao, R Chellappa, P J Phillips, and A Rosenfeld. Face recognition: A Literature Survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

Chellappa R Zhao W. and Phillips P. J. Subspace linear discriminant analysis for face recognition. Technical report, Center for Automation Research, University of Maryland, College Park, MD., 1999.