Journal of Machine Learning Research 8 (2007) 1165-1195 Submitted 8/06; Revised 1/07; Published 5/07

Local Discriminant Wavelet Packet Coordinates for Face Recognition

Chao-Chun Liu
Dao-Qing Dai[*] STSDDQ@MAIL.SYSU.EDU.CN
Center for Computer Vision and Department of Mathematics
Sun Yat-Sen (Zhongshan) University
Guangzhou, 510275 China

Hong Yan[†] H.YAN@CITYU.EDU.HK
Department of Electric Engineering
City University of Hong Kong
83 Tat Chee Avenue
Kowloon, Hong Kong, China

Editor: Donald Geman

Abstract

Face recognition is a challenging problem due to variations in pose, illumination, and expression. Techniques that can provide effective feature representation with enhanced discriminability are crucial. Wavelets have played an important role in image processing for its ability to capture localized spatial-frequency information of images. In this paper, we propose a novel *local discriminant coordinates* method based on wavelet packet for face recognition to compensate for these variations. Traditional wavelet-based methods for face recognition select or operate on the most discriminant subband, and neglect the scattered characteristic of discriminant features. The proposed method selects the most discriminant coordinates uniformly from all spatial frequency subbands to overcome the deficiency of traditional wavelet-based methods. To measure the discriminability of coordinates, a new dilation invariant entropy and a maximum *a posterior* logistic model are put forward. Moreover, a new *triangle square ratio* criterion is used to improve classification using the Euclidean distance and the cosine criterion. Experimental results show that the proposed method is robust for face recognition under variations in illumination, pose and expression.

Keywords: local discriminant coordinates, invariant entropy, logistic model, wavelet packet, face recognition, illumination, pose and expression variations

1. Introduction

Face recognition (Zhao et al., 2003; Jain et al., 2004) has become one of the most active research areas in pattern recognition. It plays an important role in many application areas, such as human-machine interaction, authentication and surveillance. However, the wide-range variations of human face, due to pose, illumination, and expression, result in a highly complex distribution and deteriorate the recognition rate. It seems impractical to collect sufficient prototype images covering all the possible variations. Therefore, how to construct a small-size-training face recognizer robust to environmental variations is a challenging research issue. Wavelets have been successfully used in image

∗. Also Department of Electric Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China. Dao-Qing Dai is the corresponding author.
†. Also School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia.

processing. Their ability to capture localized spatial-frequency information of image motivates us to use them for feature extraction. In this study, we investigate a new approach by extracting the features not sensitive to environmental changes from a wavelet packet dictionary.

Generally, feature extraction, discriminant analysis and classifying criterion are the three basic elements of a face recognition system. The performance and robustness of face recognition could be enhanced by improving these elements. Feature extraction in the sense of some linear or non-linear transform of the data with subsequent feature selection is commonly used for reducing the dimensionality of facial image so that the extracted features are as representative as possible. A lot of work on face recognition has been carried out based on similarities analysis (P. Howland and Park, 2006; Belhumeur et al., 1997; Jiang et al., 2006; Martinez and Zhu, 2005; Vaswani and Chellappa, 2006; Xiang et al., 2006; Zhao et al., 2003). A well-known feature extraction method is called FisherFace, based on linear discriminant analysis (LDA), which linearly projects the image space to a low-dimensional subspace so as to discount environmental variations (Belhumeur et al., 1997; Fukunaga, 1990). This method is a statistical linear projection method which largely relies on the representation of the training samples. On the other hand, wavelet-based methods with no special focus on the training data have been used for feature extraction (Mallat, 1989; Coifman et al., 1992). The decomposition of the data into different frequency ranges allows us to isolate the frequency components introduced by intrinsic deformations due to expression or extrinsic factors (like illumination) into certain subbands. Wavelet-based methods prune away these variable subbands, and focus on the subbands that contain the most relevant information to better represent the data. WaveletFace (Chien and Wu, 2002) only uses the low-frequency subband to present the basic figure of an image, and ignores the efficacy of high-frequency components. Our previous study (Dai and Yuen, 2006) uses a wavelet enhanced regularized discriminant analysis after dimensionality reducing with low-pass filter to solve the small sample size problem, which is also a method based on the low frequency subband. Similarly, some other studies (Feng et al., 2000; Ekenel and Sanker, 2005; Zhang et al., 2004, 2005) employ the traditional transform (e.g., ICA, PCA, Neural Networks) to enhance the discriminant power in one or several special subbands, the latter always fuse the discriminant power in these different subbands for final classification (Ekenel and Sanker, 2005; Zhang et al., 2005). Moreover, as a generalization of the wavelet transform, the wavelet packet not only offers an attractive tool for reducing the dimensionality by feature extraction, but also allows us to create localized subbands of the data in both space and frequency domains. Saito and Coifman introduced the *local discriminant basis* (LDB) algorithm based on a best-basis paradigm to search for the most discriminant subbands (basis) that illuminates the dissimilarities among classes from the wavelet packet dictionary (Coifman and Saito, 1994; Saito and Coifman, 1994, 1995). Some studies (Saito et al., 2002; Stranss et al., 2003) constructed the modified LDB later. In Kouzani et al. (1997), the best-basis algorithm of Coifman and Wicherhauser (1992) is used to search for the wavelet packet basis for face representation. In Bhagavatula and Savvides (2005), PCA is performed in wavelet packet subbands and the subbands which generalize better across illumination variations for face recognition are sought. All the methods on these studies are based on the whole discriminant subband.

It is known that a good feature extractor for a face recognition system is claimed to select as many the best discriminant features as possible, which are not sensitive to arbitrary environmental variations. Nastar and Ayach (1996) investigated the relationships between variations in facial appearance and their deformation spectrum. They found that facial expressions and small occlusions affect the intensity manifold locally. Under frequency-based representation, only high-frequency

spectrum is affected. Moreover, changes in pose or scale of a face and most illumination variations affect the intensity manifold globally, in which only their low-frequency spectrum is affected. Only a change in face will affect all frequency components (Zhang et al., 2004). So there are no special subbands whose all coordinates are not sensitive to these variations. In each subband, there may be only segmental coordinates which have enough discriminant power to distinguish different person, the remainder may be sensitive to environmental changes, but the methods based on the whole subband will also extract these sensitive features. Moreover, there may be no special subbands containing all the best discriminant features, because the features not sensitive to environmental variations are always distributed in different coordinates of different subbands locally. The methods based on the segmental subbands will lose some good discriminant features. Furthermore, in different subbands, the amount and distribution of best discriminant coordinates are always different. Many less discriminant coordinates in one subband may add up to a larger discriminability than another subband whose discriminability is added up with few best discriminant coordinates and residual small discriminant coordinates (Saito et al., 2002), then the few best discriminant coordinates will be discarded by the methods which search for the best discriminate subbands, but only the few best discriminant coordinates are needed. So the best discriminant information selection should be independent of their seated subbands, and only depends on their discriminability for face recognition. However, the methods based on the whole subband neglect the distribution of features, they are deficient to select the best discriminant features sometimes.

Moreover, how to measure the discriminability of coordinate is one crucial element of the whole algorithm. We translate it into the separability of each coordinate-loading ensemble, and propose a new dilation invariant entropy which is independent of the order of magnitude (OM), instead of deficient absolute "distance" measures. Furthermore, we construct a maximum a posterior (MAP) logistic model to produce a separability measure function which presents factually the separability of each coordinate-loading ensemble, that is, discriminability of each coordinate. Based on the new dilation invariant entropy and its derived separability measure function, any two coordinates are comparable for their discriminability, either they locate in the same subband or different subbands.

To solve the "*small sample size*" (SSS) problem, we use the complete linear discriminant analysis (CLDA) idea (Yang et al., 2005) which captures both regular and irregular discriminant information and makes a more powerful discriminator. For classifying criterion, the traditional Euclidean distance cannot measure the similarity very well when there exist illumination variations on facial images, and the cosine criterion is unsatisfactory when there exist pose and expression changes. Thus, we propose a new *triangle square ratio* criterion. Experimental results show that it can overcome the deficiency of the Euclidean distance and cosine criterion very well.

In this paper, to deal with illumination, pose and expression problems, we propose a new *local discriminant coordinates* (LDC) algorithm to select uniformly the most discriminant independent coordinates in all spatial frequency subbands for face recognition, in order to overcome the limitation of the methods based on whole subband. Experimental results show that our LDC feature extraction has almost overcome the shortcomings of the methods based on subband and improves the effect of feature extraction for face recognition under different environmental variations.

The contribution of this paper consists of the following:

- Further extension of wavelets to face recognition to deal with illumination, pose and expression problems.

- Introduction of a dilation invariant entropy and a maximum *a posterior* logistic model for selection of wavelet packet coordinates.
- Use of a new similarity criterion coupled with the nearest neighbor classifier.
- Design of a face recognition system, which solves the small sample size problem and is robust to variations in illumination, pose and expression.

The paper is organized as follows. In Section 2, the wavelet packet decomposition and the local discriminant basis algorithm will be introduced. Our proposed algorithm and the whole procedure will be presented in Section 3. In Section 4, experimental results are presented, followed by discussions and conclusion in Section 5.

2. Feature Extraction by Local Discriminant Basis

In this section, we first make a review on the wavelet packet decomposition, then the local discriminant basis (LDB) algorithm and the modified LDB algorithm are introduced.

2.1 The Wavelet Packet Decomposition

Wavelets are functions that satisfy certain mathematical requirements and are used as basis functions in representing data at different scales and time-frequency locations. Wavelets (Kouzani et al., 1997; Vaidyanathan, 1993) can be generated from a two-channel filter bank method which uses repeated filtering and downsampling to decompose signals into time-frequency subbands. The two-channel filter bank has a lowpass filter which removes the high frequencies and a highpass filter which removes the low frequencies. For the wavelet transform, only the lowpass filtered subband is further iterated. As a generalization of the wavelet transform, the two-channel filter banks are iterated over the lowpass and the highpass subbands in the wavelet packet decomposition. This generates a tree structure which provides many possible wavelet packet bases, accordingly, signals are decomposed into a time-frequency dictionary.

When dealing with images, the wavelet decomposition or the wavelet packet decomposition is first applied along the rows of the images, then their results are further decomposed along the columns. This results in four decomposed subimages L_1, H_1, V_1 and D_1. These subimages represent different frequency localizations of the original image which refer to Low-Low, Low-High, High-Low and High-High respectively. Their frequency components comprise the original frequency components but now in distinct ranges. While the process being iterated, only L_1 is further decomposed in the wavelet decomposition, but all L_1, H_1, V_1 and D_1 are further decomposed in the wavelet packet decomposition. Figure 1 shows a two-dimensional examples of a facial image for the wavelet decomposition and the wavelet packet decomposition with depth 2.

2.2 The Local Discriminant Basis (LDB) Algorithm

The *local discriminant bases* algorithm (Coifman and Saito, 1994; Saito and Coifman, 1994, 1995) uses an adjustment of dictionary, or a wavelet packet decomposition tree which offers a library of orthonormal basis localized both in space and in frequency. Before proceeding further, let us set our notations. Let $X = \{x_1, x_2, \cdots, x_N\}$ be an ensemble of training samples with K classes, $X = \bigcup_{y=1}^{K} X_y$, and $X_y = \{x_1^y, x_2^y, \cdots x_{N_y}^y\}$, where N_y is the number of samples belong to class y, and $N = \sum_{y=1}^{K} N_y$.

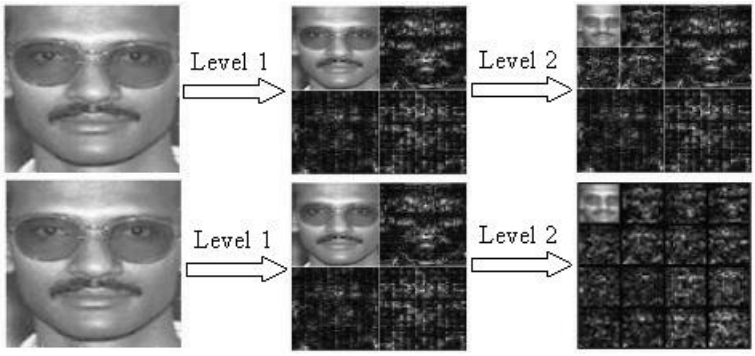

Figure 1: *(Top)* The two-dimensional wavelet decomposition of facial image with depth 2. *(Bottom)* The two-dimensional wavelet packet decomposition of facial image with depth 2.

We use $\mathcal{D}$ to represent the space-frequency dictionary consisting of a collection of wavelet packet subbands $\{B_j\}$, $j = 1, \cdots, (4^{\mathcal{L}+1}-1)/3$, where $B_j = \{b_{j1}, b_{j2}, \cdots, b_{jn_j}\}$, $b_{ji}(i = 1, 2, \cdots, n_j)$ are wavelet packet coefficients and n_j is the size of wavelet packet subband B_j, $\mathcal{L}$ is the decomposition level of wavelet packet.

The LDB algorithm first decomposes the training samples in the dictionary $\mathcal{D}$, then sample energies at the basis coordinates are accumulated for each sample class separately to form a space-frequency energy distribution per class. Let $\Gamma^{(y)}(B_j)$ be a normalized energy of class y samples presented on the subbands B_j:

$$\Gamma^{(y)}(B_j) = (\Gamma^{(y)}(b_{j1}), \Gamma^{(y)}(b_{j2}), \cdots, \Gamma^{(y)}(b_{jn_j})) \quad \forall B_j \subset \mathcal{D}, \tag{1}$$

$$\Gamma^{(y)}(b_{jt}) \triangleq \frac{\sum_{i=1}^{N_y} \left| b_{jt} \cdot x_i^y \right|^2}{\sum_{i=1}^{N_y} \left\| x_i^y \right\|^2} \tag{2}$$

where $\cdot$ denotes the standard inner product in the Euclidean space. The loss function ϕ_1 is used to measure "distances" among K vectors $\Gamma^{(1)}(B_j), \Gamma^{(2)}(B_j), \cdots, \Gamma^{(K)}(B_j)$:

$$\phi_1(B_j) = \phi_1(\Gamma^{(1)}(B_j), \Gamma^{(2)}(B_j), \cdots, \Gamma^{(K)}(B_j)) \triangleq \sum_{\substack{m,n=1 \\ m \neq n}}^{K} d^*(\Gamma^{(m)}(B_j), \Gamma^{(n)}(B_j)) \tag{3}$$

where $d^*(\cdot,\cdot)$ is a "distance" measure, it can be the l^2 distance, the relative entropy, or the J-Divergence. Then $\phi_1(B_j)$ will be a measure of efficacy of the subband B_j for classification, and local discriminant basis are selected by the best-basis algorithm (Coifman and Wicherhauser, 1992) using the following criterion:

$$\Psi = \arg\max_{B_j \in \mathcal{D}} \phi_1(B_j). \tag{4}$$

The final step is to construct traditional discriminant analysis (e.g., LDA, CT) with features derived from the LDB feature extraction.

2.3 The Modified LDB (MLDB) Algorithm

In Saito et al. (2002), a modified version of the LDB algorithm is introduced using the empirical probability distributions instead of the space-frequency energy distribution as their selection strategy to eliminate some less discriminant coordinates in each subband locally. Let

$$\delta_{jt} \stackrel{\Delta}{=} \phi_2(\Gamma^{(1)}(b_{jt}), \Gamma^{(2)}(b_{jt}), \cdots \Gamma^{(K)}(b_{jt})) = \sum_{\substack{m,n=1 \\ m \neq n}}^{K} d^*(\Gamma^{(m)}(b_{jt}), \Gamma^{(n)}(b_{jt})) \tag{5}$$

that is, the discriminability of coordinate b_{jt} $(t = 1, 2, \cdots, n_j)$. Then the measure of the discriminability of B_j is obtained by summing only the $n_0 (< n_j)$ largest terms, that is,

$$\phi_2(B_j) \stackrel{\Delta}{=} \sum_{t=1}^{n_0} \delta_{j(t)} \tag{6}$$

where $\{\delta_{j(t)}\}$ is the decreasing rearrangement of $\{\delta_{jt}\}$, and local discriminant basis are selected by the best-basis algorithm using the criterion (4) as LDB. The final step is the same as LDB.

Although the MLDB algorithm may overcome some limitations of LDB, the selection of coordinates is only limited to each subband so that coordinates in different subbands are still incomparable.

3. The General Framework of the LDC Algorithm

Our LDC algorithm uses a ternary architecture similar to LDB. We use the wavelet packet feature extraction at the first step. The main difference between LDB and our LDC algorithm is the nature of "distance" measure and feature selection strategy. We propose a new dilation invariant entropy to take the place of traditional traditional absolute "distance" measures. This ensures that the comparison of discriminability among all coordinates is independent of spatial frequency subbands. Thus, our selection can be based on all coordinates of the dictionary, but not the subbands themselves.

Moreover, LDB uses only the between-class difference, and ignores the within-class difference. This may lead to an unsatisfactory discriminability. The solution presented in this paper makes use of the maximum a posterior (MAP) logistic model. Its derived separability measure function will get a contrastive term to ensure not only the within-class difference is low, but also the between-class difference is large. Our LDC algorithm does not need the best-basis algorithm (Coifman and Wicherhauser, 1992) used in LDB, it ensures that we can select the most discriminant features without any impact of the best-basis algorithm. Subsequently, the LDC algorithm uses the complete linear discriminant analysis (CLDA) to solve the "*small sample size*" (SSS) problem, instead of the traditional LDA or CT in LDB. Finally, we modify the Euclidean distance and the cosine criterion in the nearest neighbor classifier, and replace them with the triangle square ratio criterion for classification.

3.1 The Wavelet Packet Decomposition in our LDC Algorithm

In the LDC algorithm, the wavelet packet technique is used to decompose an image into subbands that are localized in both space and frequency domains, and offers a choice of optimal coordinates for the representation of a human face. Therefore, it is possible to seek the most discriminative coordinates for classification. Because each child subband is derived from its parent subband at the above level, the coordinates in the two levels are linearly dependent. In the first experiment in

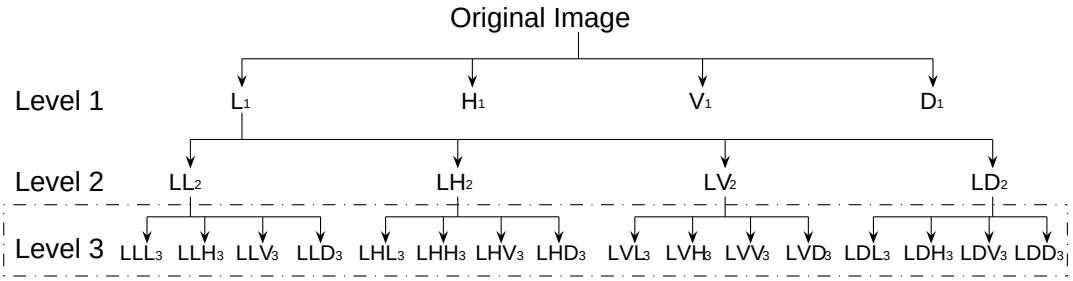

Figure 2: The wavelet decomposition tree used in this study. The *dashed part* is the spatial-frequency dictionary $\mathcal{D}$ in the LDC algorithm

Section 4, we will search for the most discriminant level by the best performance of its selected coordinates. Because it is more time-consuming when the decomposition level $\mathcal{L}$ is larger than 4, $\mathcal{L} = 4$ will be used in the experiment. In the first level, four subband images—L_1, H_1, V_1, D_1—are obtained. However, the high frequency H_1, V_1, D_1 are sensitive to noises in facial images, and Ekenel and Sanker (2005) claimed that they have low performance for classification. Moreover, the results in Table 4 show that our dilation invariant entropy used in the LDC algorithm may extract few high frequency components which may slightly affect the performance, also for the sake of computational efficiency, the H_1, V_1, D_1 components are not further decomposed. Our experimental results show that Level 3 has better performance than Level 1, 2, 4, and the same results are also presented in Chien and Wu (2002). In fact, with the further wavelet packet decomposition, more fine scale information which may have good discriminant power is generated, however, the resolution of subband images becomes lower so that less information exists for the purpose of object localization (Grewe and Brooks, 1997). Neither little scale information nor little localization information can generate a judicious combination which has best discriminate power, so Level 3 which may give a suitable tradeoff between scale information and localization information is used in some studies (Chien and Wu, 2002; Feng et al., 2000). We also use Level 3 in the LDC algorithm, and our spatial-frequency dictionary $\mathcal{D}$ consists of 16 subbands in Level 3 (a subset of the dictionary in the LDB algorithm) (see Figure 2). The *Daubechies db*4 wavelet will be used for image decomposition (Daubechies, 1990), if the sizes of facial images are not the dyadic numbers, we will apply zero-padding extension to create the smallest dyadic images for the wavelet packet decomposition.

3.2 The Dilation Invariant Entropy

In this subsection, we first point out the deficiency of absolute "distance" measures in wavelet-based methods, then introduce our dilation invariant entropy and its property.

3.2.1 DEFICIENCY OF ABSOLUTE "DISTANCE" MEASURES IN WAVELET-BASED METHODS

To introduce our new dilation invariant entropy, we first list several traditional discriminant measures. Given two nonnegative sequences $w = (w_1, w_2, \cdots, w_n)$, $z = (z_1, z_2, \cdots, z_n)$, the *square* ℓ^2-

norm is defined by

$$d^{s\ell^2}(w,z) \stackrel{\Delta}{=} \|w-z\|_2^2 = \sum_{i=1}^{n}(w_i-z_i)^2. \tag{7}$$

Suppose $\sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} z_i = 1$, then the *Kullback-Leibler divergence* (Kullback and Leibler, 1951), also known as *relative entropy*, is defined by

$$d^{KLD}(w,z) \stackrel{\Delta}{=} \sum_{i=1}^{n} w_i \log \frac{w_i}{z_i} \tag{8}$$

with the convention that $\log 0 = -\infty$, $\log \gamma/0 = \infty$ for $\gamma > 0$ and $0(\pm\infty) = 0$. A symmetric version of d^{KLD} is the *J-Divergence* (Kullback and Leibler, 1951) given by

$$d^{JDIV}(w,z) \stackrel{\Delta}{=} \frac{d^{KLD}(w,z) + d^{KLD}(z,w)}{2}. \tag{9}$$

It is easy to show that measures in Equations (7)-(9) are *additive discriminant measure*, that is,

$$d^*(w,z) = \sum_{i=1}^{n} d^*(w_i,z_i) \quad (*=s\ell^2, KLD, JDIV). \tag{10}$$

From Equations (3) and (10), we know that the discriminant measure of subband B_j in the LDB algorithm can be written as

$$\phi_1(B_j) = \sum_{\substack{m,n=1 \\ m\neq n}}^{K} \sum_{t=1}^{n_j} d^*(\Gamma^{(m)}(b_{jt}), \Gamma^{(n)}(b_{jt})). \tag{11}$$

Also from Equations (5) and (6), we know that the discriminant measure of subband B_j in the MLDB algorithm can be written as

$$\phi_2(B_j) = \sum_{\substack{m,n=1 \\ m\neq n}}^{K} \sum_{t=1}^{n_0} d^*(\Gamma^{(m)}(b_{j(t)}), \Gamma^{(n)}(b_{j(t)})), \ (n_0 < n_j) \tag{12}$$

where $b_{j(t)}, \ (t = 1, \cdots, n_0)$ are the first n_0 coordinates with largest discriminability in subband B_j.

However, there are no normalized conditions imposed in each subband when the decomposition level $\mathcal{L} > 0$, because for each subband $B_j (j > 1)$ (B_1 is the original image)

$$\sum_{t=1}^{n_j} \Gamma^{(y)}(b_{jt}), \sum_{t=1}^{n_0} \Gamma^{(y)}(b_{j(t)}) < 1 \ \text{ and } \ \sum_{t=1}^{n_j} \Gamma^{(y)}(b_{jt}) \not\equiv C_0, \sum_{t=1}^{n_0} \Gamma^{(y)}(b_{j(t)}) \not\equiv C_1 \quad \forall y$$

where C_0, C_1 are constants independent of y and B_j. Without the normalized conditions, the absolute "distance" measures (7)-(9) will lead to a jeopardy that $\phi_1(B_j)$ and $\phi_2(B_j)$ depend absolutely on the order of magnitude (OM) of $\Gamma^{(m)}(b_{jt})$, $\Gamma^{(n)}(b_{jt})$ and $\Gamma^{(m)}(b_{j(t)})$, $\Gamma^{(n)}(b_{j(t)})$ respectively.

Unfortunately, we find that the OM of coordinate loadings make much difference between lower spatial frequency subbands and higher spatial frequency subbands. For example, the coordinate loadings in the first spatial frequency subband B_6 (=LL_2 in Figure 2) of the second level may vary from 0.1 to 10, and the coordinate loadings in the second spatial frequency subband B_7 (=LH_2 in

B_6		B_7	
$\Gamma^{(1)}(b_{6t})$	$\Gamma^{(2)}(b_{6t})$	$\Gamma^{(1)}(b_{7t})$	$\Gamma^{(2)}(b_{7t})$
1.47e-04	3.86e-04	6.72e-08	4.90e-07
8.04e-05	3.57e-04	4.13e-07	5.04e-06
2.07e-04	3.91e-04	1.32e-06	1.10e-07
1.06e-04	3.57e-04	5.45e-07	7.02e-08
1.07e-04	3.83e-04	1.43e-07	5.32e-08
2.05e-04	3.57e-04	5.20e-09	7.14e-08

Table 1: Values of $\Gamma^{(y)}(b_{6t}), \Gamma^{(y)}(b_{7t})$ $(y = 1,2;\ t = 1,\cdots,6)$

Figure 2) of the second level may vary from 0.001 to 0.01. Table 1 lists an example of some values of $\Gamma^{(y)}(b_{6t}), \Gamma^{(y)}(b_{7t})$ computed by Equation (2) in latter experiment.

From Equations (11),(12) and (2), we can deduce a bad result that the coordinates in lower spatial frequency subbands have more discriminability because of the larger OM of theirs loadings, and the coordinates in higher spatial frequency subbands have less discriminability because of the smaller OM of theirs loadings. So the low spatial frequency subbands are dominant in the LDB and MLDB algorithm. However, it is unreasonable to neglect the middle and high spatial frequency components merely for small OM of their loadings. Our experimental results also show that not only low spatial frequency components, but also middle spatial frequency components are useful for face recognition.

3.2.2 THE DILATION INVARIANT ENTROPY AND ITS PROPERTY

First, we define that the separability of sample ensemble X is the probability of classifying all samples into their genuine classes by certain discriminant functions. It is well-known that the separability of X does not depend on the absolute distances based on the OM of sample values, but depends on the relative distances among all the samples in X. For each coordinate c, the coordinate loadings from all the training samples can induce a sample ensemble X^c in $\mathbb{R}^1$, and the discriminability of c is equivalent to the separability of X^c, so it is independent of the OM of coordinate loadings, and only depends on the relative distances among the coordinate loadings from all the training samples. Obviously, the "distance" measures used in LDB do not take this fact into account. So we propose a new "distance" measure derived from the J-Divergence. We call it the *dilation invariant entropy* :

$$d^{DIE}(w,z) \stackrel{\Delta}{=} \sum_{i=1}^{n} \frac{1}{2(w_i+z_i)}\left(\frac{w_i}{z_i}\log w_i + \frac{z_i}{w_i}\log z_i\right) = \sum_{i=1}^{n} \frac{(w_i-z_i)(\log w_i - \log z_i)}{2(w_i+z_i)} \tag{13}$$

where $w = (w_1, w_2, \cdots, w_n)$, $z = (z_1, z_2, \cdots, z_n)$ are two nonnegative sequences, with the convention that $\log 0 = -\infty$, $\log \gamma/0 = \infty$ for $\gamma > 0$ and $0(\pm\infty) = 0$.

Proposition 1 *The new relative entropy defined by Equation (13) is dilation invariant.*

Proof Suppose the dilation transform $f : w \in \mathbb{R}^n \to f(w) = aw \in \mathbb{R}^n$, $a(>0)$ is a dilation constant, then

$$
\begin{aligned}
d^{DIE}(f(w),f(z)) = d^{DIE}(aw,az) &= \sum_{i=1}^{n} \frac{(aw_i - az_i)(\log aw_i - \log az_i)}{2(aw_i + az_i)} \\
&= \sum_{i=1}^{n} \frac{a(w_i - z_i)(\log w_i + \log a - \log z_i - \log a)}{2a(w_i + z_i)} \\
&= \sum_{i=1}^{n} \frac{(w_i - z_i)(\log w_i - \log z_i)}{2(w_i + z_i)} \\
&= d^{DIE}(w,z).
\end{aligned}
$$

■

In fact, the new dilation invariant entropy is the generalization of the J-Divergence (when the J-Divergence satisfies the constraint: $w+z=1$) because

$$
\begin{aligned}
d^{DIE}(w,z) &= \sum_{i=1}^{n} \frac{(w_i - z_i)(\log w_i - \log z_i)}{2(w_i + z_i)} \\
&= \sum_{i=1}^{n} \frac{1}{2}\left(\frac{w_i}{w_i + z_i} - \frac{z_i}{w_i + z_i}\right)\left(\log \frac{w_i}{w_i + z_i} - \log \frac{z_i}{w_i + z_i}\right) \\
&= \frac{1}{2}\left(\sum_{i=1}^{n} w_i' \log \frac{w_i'}{z_i'} + \sum_{i=1}^{n} z_i' \log \frac{z_i'}{w_i'}\right) \\
&= d^{JDIV}(w',z')
\end{aligned}
\tag{14}
$$

where $w_i' = \frac{w_i}{w_i+z_i}, z_i' = \frac{z_i}{w_i+z_i}$, $w_i' + z_i' = 1$ and $w' = (w_1', \cdots, w_n'), z' = (z_1', \cdots, z_n')$. Equation (14) shows that the dilation invariant entropy normalizes the sample ensembles into unit sample ensembles with the sum formalism, so different sample ensembles are comparable for their separability. Similarly, we can define the dilation invariant ℓ^2 norm as

$$
d^{DI\ell^2}(w,z) = \left(\sum_{i=1}^{n} \left(\frac{w_i}{w_i + z_i} - \frac{z_i}{w_i + z_i}\right)^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} (w_i' - z_i')^2\right)^{\frac{1}{2}} = d^{\ell^2}(w',z'). \tag{15}
$$

In Section 4, we will conduct an experiment to test the performance of the LDC algorithm using both dilation invariant entropy and dilation invariant ℓ^2 norm.

The dilation-invariance of the new relative entropy ensures that the separability of each coordinate-loadings ensemble X^c is independent of its OM. Accordingly, the discriminability of each coordinate c can be independent of its corresponding subband. It offers a benefit that any two coordinates in the dictionary $\mathcal{D}$ are comparable for their discriminability. So all the coordinates in the dictionary can be uniformly selected by a criterion.

3.3 Feature Selection Criterion

Sometimes, maximizing the between-class difference or minimizing the within-class difference alone leads to a bad result. So in the LDC algorithm, our separability measure function will contain a term to maximize the between-class difference and minimize the within-class difference simultaneously.

3.3.1 THE MAXIMUM *a posteriori* (MAP) LOGISTIC MODEL

To select the most discriminant coordinates, we make use of the Bayesian algorithm based on minimizing the error on the training set. The Bayesian algorithm adopts a probabilistic measure of similarity based on a Bayesian MAP analysis of face differences. In the traditional methods (Wang et al., 2006; Chou, 2000), the similarity measure is used to characterize what kind of image variation is typical for the same person and what is for different persons. In this paper, the MAP similarity measure is used to choose the coordinates that make the training data set with known class labels having the minimum error. In this way the selected coordinates can make the known classification of training data set the most probable:

$$\hat{c} = \arg\max_{c} \sum_{y=1}^{K} \frac{1}{\#(X_y)} \int_X P_c(X_y|x) 1(x \in X_y) dP(x)$$

where $\#(\cdot)$ is the cardinal number, and $1(\cdot)$ is the indicator function. The posterior probability $P_c(X_y|x)$ can be rewritten as

$$P_c(X_y|x) = P_c(x|X_y)P_c(X_y)/P_c(x) = P_c(x|X_y)P_c(X_y)/P(x).$$

Since $P(x)$ is not a function of the class index and thus has no effect in the MAP decision, the needed probabilistic knowledge can be represented by the class prior distribution $P_c(X_y)$ and the conditional probability $P_c(x|X_y)$ which will be modeled by logistic functions.

Definition 1 *The prior distribution $P_c(X_y)$ is defined as*

$$\begin{aligned} P_c(X_y) &= \frac{1}{T} \frac{1}{1+\exp(-d^{DIE}(\Gamma^{(y)}(c), \Gamma^{(0)}(c)))}, \\ T &= \sum_{y=1}^{K} \frac{1}{1+\exp(-d^{DIE}(\Gamma^{(y)}(c), \Gamma^{(0)}(c)))}. \end{aligned} \tag{16}$$

Definition 2 *The conditional probability $P_c(x_i^y|X_y)(x = x_i^y)$ is defined as*

$$P_c(x_i^y|X_y) = \frac{1}{1+\exp(d^{DIE}(\Gamma_i^{(y)}(c), \Gamma^{(y)}(c))} \tag{17}$$

where $\Gamma^{(y)}(c)$ is defined by Equation (2), representing the normalized spatial-frequency energy map of class y on coordinate c, and can be thought of as the center of class y. Similar to that of LDB, we set

$$\Gamma_i^{(y)}(c) \triangleq \frac{|c \cdot x_i^y|^2}{\|x_i^y\|^2}, \qquad \Gamma^{(0)}(c) \triangleq \frac{\sum_{y=1}^{K} \sum_{i=1}^{N_y} |c \cdot x_i^y|^2}{\sum_{y=1}^{K} \sum_{i=1}^{N_y} \|x_i^y\|^2}. \tag{18}$$

$\Gamma_i^{(y)}(c)$ represents the normalized spatial-frequency energy map of sample x_i^y on coordinate c, and $\Gamma^{(0)}(c)$ represents the normalized spatial-frequency energy map of all the training samples on coordinate c, which can be considered as the center of all samples.

The properties of the probability functions $P_c(X_y)$ and $P_c(x_i^y|X_y)$ can be made clear by considering the sigmoid function:

$$f(d) = \frac{1}{1+\exp(-\gamma d + \theta)}$$

with θ normally set to zero and γ set to 1 for $P_c(X_y)$ and -1 for $P_c(x_i^y|X_y)$. When $\gamma = 1$, $f(d)$ is a monotonically increasing function, a larger $d^{DIE}(\Gamma^{(y)}(c), \Gamma^{(0)}(c)))$ means that it is more probable to separate class set X_y. Contrarily, when $\gamma = -1$, $f(d)$ is a monotonically decreasing function, a smaller $d^{DIE}(\Gamma_i^{(y)}(c), \Gamma^{(y)}(c))$ means that sample x_i^y is more likely to belong to class set X_y. In fact, the idea of MAP logistic model is derived from the Fisher criterion. Moreover, the sigmoid function can effectively allay the effect of outliers which have great effect on the Fisher criterion.

3.3.2 SEPARABILITY MEASURE AND FEATURE SELECTION CRITERION

For the given training data set, the empirical probability measure $P(x)$ defined on the training data set is a discrete probability measure that assigns equal mass at each sample. We define the separability measure as

$$\begin{aligned} SM(c) &= \sum_{y=1}^{K} \frac{1}{\#(X_y)} \int_X P_c(X_y|x) 1(x \in X_y) dP(x) \\ &\approx \sum_{y=1}^{K} \frac{1}{N_y} \sum_{i=1}^{N_y} P_c(x_i^y|X_y) P_c(X_y) \\ &= \sum_{y=1}^{K} P_c(X_y) \left(\frac{1}{N_y} \sum_{i=1}^{N_y} P_c(x_i^y|X_y) \right). \end{aligned} \tag{19}$$

In fact, the separability measure defined by Equation (19) is an empirical measure. If the training samples are obtained by an independent sampling from a space with a fixed probability distribution $P_0(x)$, the empirical probability distribution $P(x)$ will converge to $P_0(x)$ in distribution as $N \to \infty$. Then the empirical measure defined on the N independent training samples will converge to the expected measure as the sample size N increases. With sufficient training samples, the empirical measure is an estimate of the expected measure. The goodness of this estimate is determined by the training sample size N and the convergence rate of the empirical probability measure $P(x)$ to the limit distribution $P_0(x)$.

Furthermore, we use the following criterion for feature selection:

Criterion: Select uniformly the first N_0 coordinates from the dictionary $\mathcal{D}$ with largest separability measure defined by Equation (19).

3.4 Discriminant Analysis

LDA (Fukunaga, 1990) is a linear statistic classification method, which tries to find a linear transform so that after its application the scatter of sample vectors is minimized within each class and the scatter of mean vectors around the total mean vector is maximized simultaneously.

Let the between-class scatter operator S_b and the within-class scatter operator S_w be:

$$S_b = \frac{1}{N} \sum_{y=1}^{K} N_y (m_y - m_0)(m_y - m_0)^T, \; S_w = \frac{1}{N} \sum_{y=1}^{K} \sum_{i=1}^{N_y} (x_i^y - m_y)(x_i^y - m_y)^T$$

where m_y is the mean of the mapped training sample of class y, and m_0 is the mean across all the mapped training samples. Then the Fisher criterion function can be defined by

$$J_1(\varphi_1) = \frac{\varphi_1^T S_b \varphi_1}{\varphi_1^T S_w \varphi_1}, \quad (\varphi_1 \neq 0, \|\varphi_1\| = 1). \tag{20}$$

The solution to maximizing $J_1(\varphi_1)$ can be found by searching for a direction which maximizes the projected class means (the numerator) while minimizing the class variances in this direction (the denominator).

However, the LDA algorithm often suffers from the "*small sample size*" (SSS) problem which exists in high-dimensional pattern recognition tasks, where the number of available samples is smaller than the dimensionality of the samples. Many methods (Mika et al., 1999; Baudat and Anouar, 2000; S. Mika and Müller, 2003; Yang, 2002) discard the discriminant information contained in the null space of S_w. But a significant result is a finding that there exists crucial discriminative information in the null space of S_w (Chen et al., 2000; Zhuang and Dai, 2007; Yang and Yang, 2003; Yu and Yang, 2001). We proposed the use of regularization (Dai and Yuen, 2003), but it involves a determination of parameters. Yang et al. (2005) proposed a complete kernel Fisher discriminant analysis algorithm which makes full use of two kinds of discriminant information, regular and irregular in kernel feature space. Its advantage is that no estimation of parameter is needed. Based on their idea, we use complete linear discriminant analysis (CLDA) in the LDC algorithm to solve the SSS problem.

In Equation (20), if the within-class scatter operator S_w is invertible, $\varphi_1^T S_w \varphi_1 > 0$ always holds for every nonzero vector φ_1, and the Fisher criterion can be directly employed to extract a set of optimal discriminant vectors. If S_w is singular, there always exist vectors satisfying $\tilde{\varphi}^T S_w \tilde{\varphi} = 0$. These vectors are from the null space of S_w ($null(S_w)$) and can be very effective if they satisfy $\tilde{\varphi}^T S_b \tilde{\varphi} > 0$ at the same time (Chen et al., 2000; Zhuang and Dai, 2007; Yang and Yang, 2003; Yu and Yang, 2001). In this case, the Fisher criterion degenerates into the following between-class scatter criterion:

$$J_2(\varphi_2) = \varphi_2^T S_b \varphi_2, \quad (\|\varphi_2\| = 1). \tag{21}$$

CLDA uses the between-class scatter criterion defined in Equation (21) to derive the irregular discriminant vectors from $null(S_w)$, while using the standard Fisher criterion defined in Equation (20) to derive the regular discriminant vectors from $range(S_w)$.

In our experiments, we capture all the regular discriminant vectors which satisfy $J_1(\varphi_1) > 0$ from the range space of S_w, simultaneously, we capture all the irregular discriminant vectors which satisfy $J_2(\varphi_2) > 0$ from the null space of S_w.

3.5 A New Criterion for the Nearest Neighbor (NN) Classifier

After the discriminant features are extracted, a remaining key element of face recognition is to design a robust classifier. Because there are large numbers of classes in face recognition problems, we do not use the hyperplane classifier, but the NN classifier which is more suitable for such many-class problems. Because the NN classifier forms class boundaries with piecewise linear hyperplanes, any classifying border can be approximated by a series of hyperplanes defined locally. Moreover, classifying criterion is the core of the design. The Euclidean distance is the most popular one which exhibits the distance between two vectors intuitively. However, it ignores the correlation which is also important for measuring the similarity of two vectors. On the contrary, the cosine criterion (Zhang and Korfhage, 1999) exhibits the correlation, but ignores the distance between two vectors. In order to improve the Euclidean distance and cosine criterion, we propose a new *triangle square*

ratio which takes into account of both distance and correlation between two vectors. Suppose υ_1, υ_2 are two vectors, the *triangle square ratio* is defined as

$$TSR(\upsilon_1,\upsilon_2)=\frac{\|\upsilon_1-\upsilon_2\|_2^2}{\|\upsilon_1\|_2^2+\|\upsilon_2\|_2^2}.$$

The triangle square ratio is a similarity measure based on the argument and modulus of each vector, as shown in proposition 2.

Proposition 2 *Suppose* θ *is the include angle of* υ_1 *and* υ_2*, then* $TSR(\upsilon_1,\upsilon_2)\to 0$ *if and only if* $\|\upsilon_1\|_2\to\|\upsilon_2\|_2$ *and* $\theta\to 0$*, which implies the correlation between* υ_1 *and* υ_2 *should approach* 1.

Proof

$$\begin{aligned} TSR(\upsilon_1,\upsilon_2) &= 1-\frac{2\|\upsilon_1\|_2\cdot\|\upsilon_2\|_2}{\|\upsilon_1\|_2^2+\|\upsilon_2\|_2^2}\cos\theta \quad \text{(by the cosine law)} \\ &\geq 1-\cos\theta \quad (\text{“=” holds if and only if } \|\upsilon_1\|_2=\|\upsilon_2\|_2) \\ &\geq 0 \quad (\text{“=” holds if and only if } \theta=0). \end{aligned} \tag{22}$$

■

In fact, if υ_1 and υ_2 are unit vectors, the triangle square ratio is equivalent to Euclidean distance. Also, Equation (22) shows that triangle square ratio is a modification of cosine criterion by the term $\frac{2\|\upsilon_1\|_2\cdot\|\upsilon_2\|_2}{\|\upsilon_1\|_2^2+\|\upsilon_2\|_2^2}$. If $\|\upsilon_1\|_2=\|\upsilon_2\|_2$, it is equivalent to *cosine* criterion. Moreover, we have done large numbers of numerical experiments which exclusively show that $\sqrt{TSR(\upsilon_1,\upsilon_2)}$ satisfies the triangle inequality, and the symmetric and positive definitive properties are obvious. So we guess $\sqrt{TSR(\upsilon_1,\upsilon_2)}$ is a distance measure, its proof in theory is an open problem.

Experimental results in Section 4 show that the triangular square ratio is more robust against illumination variations than the Euclidean distance, whilst retaining the robustness against pose and expression changes as the Euclidean distance. On the other hand, although the triangular square ratio marginally underperforms the cosine criterion when there are variations of illumination, it can obviously outperform the cosine criterion when there are changes of pose and expression.

3.6 The Procedure of Proposed LDC Algorithm

We summarize our local discriminant wavelet packet coordinates algorithm as follows:
Step 1: The wavelet packet transform

Expand each training sample x_i^y into the spatial-frequency dictionary $\mathcal{D}$ (see Figure 2) by the wavelet packet decomposition, then x_i^y will be represented by the loadings of coordinates in $\mathcal{D}$.
Step 2: The LDC selection transform

(2.a) For each coordinate c in the dictionary $\mathcal{D}$, use the Equations (2), (18), (16), and (17) to compute its prior distribution $P_c(X_y)$ and conditional probability $P_c(x_i^y|X_y)$, whereafter, compute its separability measure defined by Equation (19), that is, its discriminability.

(2.b) Select the first N_0 coordinates from $\mathcal{D}$ with the largest discriminability. Whereupon, each training sample x_i^y can be represented by a feature vector v_i^y which is formed by the loadings of the selected coordinates. These feature vectors form a new feature space $\mathcal{F}$.
Step 3: The CLDA transform

Use the Equations (20) and (21) to construct the subspace template by the complete linear discriminant analysis (CLDA) in $\mathcal{F}$.

Step 4: Testing a new probe sample

For a new probe sample, it will be expanded into the spatial-frequency dictionary $\mathcal{D}$ by the wavelet packet decomposition, and be extracted the loadings of the corresponding coordinates selected in Step 2 to form a new feature vector v_{new}. Then v_{new} will be projected to the subspace constructed in Step 3 and classified by the nearest neighbor classifier.

Simply, the LDC algorithm can be represented as:

$$Output = T_3 \cdot T_2 \cdot T_1 \cdot Input$$

where T_1 is the wavelet packet transform, T_2 is the LDC selection transform. and T_3 is the CLDA transform.

3.7 Computational Complexity Comparison of the LDC and LDB Algorithm

The framework of the LDC algorithm is similar to LDB. We compare their computational complexity step by step:

Step 1: The wavelet packet transform

The same procedure of the LDC and LDB algorithms cost $O(N \cdot n_r n_c \cdot \mathcal{L})$, where $n_r \times n_c$ is the size of facial images, $\mathcal{L}$ is the level of the wavelet packet decomposition.

Step 2: The LDC/LDB selection transform

(2.a) For each coordinate in $\mathcal{D}_{LDC}$, the LDC algorithm needs to compute the prior distribution (16), the conditional probability (17) and its discriminability (19), so the costs of all the coordinates in $\mathcal{D}_{LDC}$ are $O(N \cdot n_r n_c) + O(K \cdot n_r n_c) + O(N \cdot n_r n_c)$. For each subband in $\mathcal{D}_{LDB}$, the LDB algorithm needs to compute the space-frequency energy distribution (2), (1) and its measure of efficacy (3). So the costs of all the subbands in $\mathcal{D}_{LDB}$ are $O(N \cdot n_r n_c \cdot \mathcal{L}) + O(K^2 \cdot n_r n_c \cdot \mathcal{L})$.

(2.b) The LDC algorithm needs to sort all the coordinates in $\mathcal{D}_{LDC}$ by their discriminability, which costs $O(n_r n_c \cdot \log_2(n_r n_c))$. The LDB algorithm needs to select the local discriminant basis from $\mathcal{D}_{LDB}$ using the best-basis algorithm, which costs $O(\mathcal{L} \cdot 4^{\mathcal{L}})$. Then both algorithms need to represent all the training samples by new feature vectors, which cost $O(N \cdot N_0)$.

Step 3: The CLDA/LDA transform

CLDA in the LDC algorithm has the same computational complexity $O((N_0)^3)$ as LDA in the LDB algorithm in the new feature space $\mathcal{F}$.

Step 4: Testing a new probe sample

A new probe sample should be transformed by T_1, T_2, T_3 with complexity $O(n_r n_c \cdot \mathcal{L}) + O(N_0) + O(N_0 \cdot N_{ev})$ and classified by the NN classifier with complexity $O(N \cdot N_{ev})$, where N_{ev} is the number of eigenvectors extracted by CLDA or LDA.

The computational complexity of Step 2 shows that the LDC algorithm is more efficient than LDB in many real applications when K is large. Table 11 also validates the fact.

4. Experiment Results

The results presented in this section are divided into five parts. First, we construct a dictionary $\mathcal{D}$ and choose a preferable N_0 for our LDC algorithm. As aforementioned, the LDC algorithm consists

of the LDC based feature extraction, the complete linear discriminant analysis (CLDA) and the nearest neighbor (NN) classifier with the triangle square ratio (TSR) criterion. In the second and third parts, we evaluate the efficacy of the LDC feature extraction and the new classifying criterion respectively. The fourth part gives the performance of the whole LDC algorithm. Some further researches of the LDC algorithm are shown in the final part.

4.1 Database

1) FERET Database: The FERET database, distributed by the National Institute of Standards and Technology, consists of 14051 eight-bit grayscale images of human heads with different expressions, poses, occlusion and illuminations (Phillips et al., 2000). Two data sets of the database are used in our experiments, one is a small data set, which contains 432 images of 72 people and each individual has six images, the other is a large data set with 255 individuals, and each person has four frontal images, the datas are extracted from four different sets, namely, Fa, Fb, Fc, and duplicate (Phillips et al., 2000). There are 1020 images in this data set. All the images are aligned by the centers of eyes and mouth, and then normalized with the resolution 92×112. Some images from both data sets of the FERET database are shown in Figure 3.

2) ORL Database: The Olivetti-Oracle Research Lab (ORL) database has 40 subjects and each subject has 10 different facial views representing various expressions, small occlusion (by glasses), different scales and orientations. So there are totally 400 facial images in the database. Each image has 92×112 pixels in gray scale. Some samples are shown in Figure 4.

3) Hybrid Database: As aforementioned, the variations of the ORL database and the FERET database are very different, which lead to unequal covariance distribution. So we blend the small FERET data set and the ORL database together, in order to test the performance of the LDC algorithm when facial images have larger illumination variations and pose, expression changes (Loog and Duin, 2004). The hybrid database has 832 images of 112 persons.

4) CMU PIE Database: The CMU Pose, Illumination, and Expression (PIE) (Sim et al., 2003) database consists of 41368 images of 68 people. Each person has images captured under 13 different poses and 43 different illumination conditions and with four different expressions. In this paper, we use a subset that focuses on illumination variations with pose and expression variations in frontal

Figure 3: Facial images of the FERET database. *(Top)* A person from the small data set. *(Bottom)* A person from the large data set.

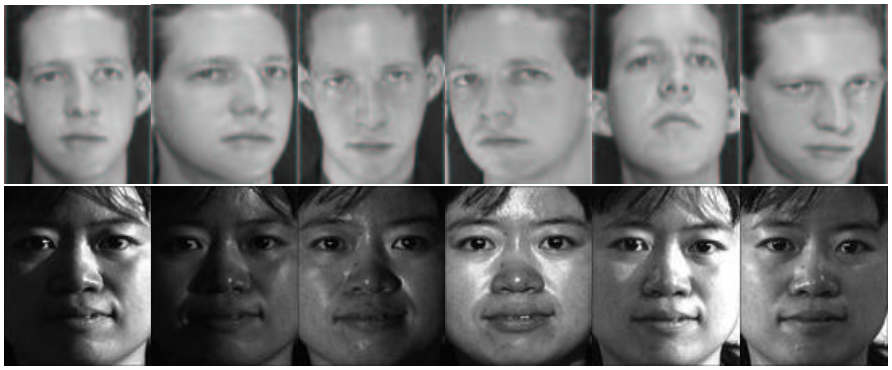

Figure 4: Segmental facial images of one person. *(Top)* From the ORL database. *(Bottom)* From the CMU PIE database.

Database	Total number of images	Number of images per person	Number of classes
ORL	400	10	40
SmallFERET	432	6	72
LargeFERET	1020	4	255
Hybrid	832	6 or 10	112
CMU-lights	2924	43	68

Table 2: Statistics for face images

view. There are 68 persons with each 43 images yielding a total of 2924 images. Each image has 92×112 pixels in gray scale. Some samples are shown in Figure 4.

The statistics of each data set is listed in Table 2.

4.2 Parameter Setting

In order to show more comparability with the PCA+CLDA, WaveletFace, LDB and MLDB algorithms and present the performance of our LDC algorithm more accurately, CLDA is used to capture the complete discriminant features in the five algorithms. The number of discriminant vectors is obtained in the same way as the LDC algorithm (see Subsection 3.4).

The FisherFace technique uses the classical PCA+LDA. The $N_{train} - K - \lambda$ (λ is the critical value which ensures S_w is non-singular) eigenvectors with largest eigenvalues are preserved on 'PCA step' (Belhumeur et al., 1997). For the PCA+CLDA algorithm, we select the first $min(N_0, M_0)$(M_0 is the number of non-zero eigenvalues) eigenvectors with largest eigenvalues on 'PCA step'. For direct LDA (DLDA), we use all the eigenvectors in their Step 2 (Yu and Yang, 2001).

The third-level lowest frequency subband LLL_3 with a matrix of $(n_r/8) \times (n_c/8)$ (where $n_r \times n_c$ is the resolution of original image) is referred to as WaveletFace (Chien and Wu, 2002). Because the LDB algorithm selects a best discriminant subset of the whole basis, we choose four subbands, and the number of selected coordinates is closest to N_0. For the MLDB algorithm, we choose N_0 coordinates as LDC on the scheme—five subbands with each 260 coordinates.

Methods	Parameters	
	number of features for discriminant analysis	classifier (criterion)
FisherFace	$N_{train} - K - \lambda$	NN(l^2)
PCA+CLDA	$min(N_0, M_0)$	NN(l^2)
DLDA	all eigenvectors	NN(l^2)
WaveletFace	subband LLL_3 in $\mathcal{D}$	NN(l^2)
LDB	four subbands	NN(l^2)
MLDB	5×260	NN(l^2)
LDC	N_0	NN(TSR)

Table 3: Parameters of aforementioned methods

In the 'Decision Step', we use the nearest neighbor classifier with the Euclidean distance for the aforementioned methods as their original forms. For our LDC algorithm, we use the new *triangle square ratio* criterion, the Euclidean distance and the cosine criterion are used for comparison. Most parameters are listed in Table 3.

The recognition rate is calculated as the ratio of the number of successful recognition and the total number of test samples. All the experiments are repeated 30 times, and the final recognition rate is the average value of the thirty results. Suppose M is the number of facial images for each person. On each database, we randomly select $i_0 (< M)$ images from each person for training, while the rest $M - i_0$ images of each individual are selected for testing. i_0 is a small integer, in order to show the performance of the LDC algorithm when there are small-size-training samples.

4.3 Construction of the Dictionary $\mathcal{D}$ and Choice of N_0

In this subsection, we conduct two experiments to construct the dictionary $\mathcal{D}$ and select a suitable N_0 for the LDC algorithm.

4.3.1 CONSTRUCTION OF THE DICTIONARY $\mathcal{D}$

In the first experiment, to construct our dictionary $\mathcal{D}$, we search for the most discriminant level by the best performance of its selected coordinates. Because it is more time-consuming when the decomposition level $\mathcal{L}$ is larger than 4, $\mathcal{L} = 4$ is used in the experiment. In order to test the effect of high frequency components and show the tolerance of the dilation invariant entropy in LDC to noise, we design two schemes: Scheme 1 uses all the subbands in the wavelet decomposition tree, Scheme 2 only uses the left subtree whose root node is L_1, that is the H_1, V_1, D_1 components are not further decomposed. For each level, we select the first 1000 coordinates by the criterion in Subsection 3.3 for both schemes. Their performances on the small FERET data set and the ORL database are shown in Table 4.

Table 4 shows that the performance of Scheme 1 is marginally underperform Scheme 2, and Scheme 1 with all the subbands in the wavelet packet tree is more time-consuming than Scheme 2 which only uses the left subtree whose root node is L_1. So Scheme 2 is adopted in the LDC algorithm. However, the effect of high frequency components is very slight, especially in the first three levels, which implies that our dilation invariant entropy has good tolerance to noise. Table 4 also shows that the third level has the best performance, it can offer a judicious combination of

Database	Training samples (i_0)	Schemes	Level			
			1	2	3	4
ORL	3	1	82.43± 2.40	91.38 ± 2.57	92.99 ± 1.60	90.43 ± 2.38
		2	82.35± 2.54	92.19 ± 2.32	93.30 ± 1.85	92.34 ± 1.78
	4	1	85.54± 2.76	94.00 ± 1.56	94.47 ± 1.68	91.79 ± 2.08
		2	85.56± 2.87	94.43 ± 1.62	95.26 ± 1.46	94.18 ± 1.35
	5	1	87.70± 2.03	95.32 ± 1.45	95.42 ± 1.71	92.57 ± 1.62
		2	87.82± 2.12	96.03 ± 1.49	96.23 ± 1.55	94.12 ± 1.59
FERET (small)	3	1	92.42± 1.93	91.88 ± 1.69	92.33 ± 2.51	92.41 ± 2.42
		2	92.42± 1.93	91.88 ± 1.77	92.30 ± 2.43	92.53 ± 2.39
	4	1	94.93± 1.40	94.79 ± 2.25	95.23 ± 2.40	94.31 ± 2.58
		2	94.93± 1.40	94.72 ± 2.04	95.60 ± 2.20	95.30 ± 2.94
	5	1	95.93± 1.41	96.30 ± 1.76	97.04 ± 1.92	96.39 ± 2.95
		2	95.93± 1.41	96.07 ± 1.90	96.90 ± 2.33	97.13 ± 2.13

$(*)\pm(**)$: $(*)$ represents the recognition rate (%), $(**)$ represents standard deviation (%).

Table 4: Effect of high frequency components and performances of level 1,2,3,4

scale information and localization information. So Level 3 is used in the LDC algorithm, and our spatial-frequency dictionary $\mathcal{D}$ consists of the first 16 subbands in the third level (see Figure 2).

4.3.2 CHOICE OF N_0

Because all of the top N_0 coordinates are used for the classification, a natural way to determine the best N_0 is to select the top N_0 coordinates and compute the average recognition rates for various different N_0. Based on the idea, we use a global to local search strategy (Müller et al., 2001) on both data sets of the FERET database. Because the computational complexity of CLDA is $O((N_0)^3)$, for the sake of computational efficiency, we set the range [100,2500] as the original wide range of N_0. In the "global" stage, we compare the performances of the LDC algorithm using the top N_0 coordinates when N_0 increases from 100 to 2500 with interval 100 on the small data set, as shown in Figure 5 *(Left)*. It shows that the LDC algorithm has a good and stable performance after $N_0 = 700$ because more good discriminant features are used. After $N_0 = 1700$, the performance slightly decreases due to the more redundant information included. So we ascertain a more precise subrange [700, 1700] where the optimal N_0 might exist.

In the "local" stage, we compare the performances of different N_0 between 700 and 1700 with interval 100 on the large data set, as shown in Figure 5 *(Right)*. From the overall comparison of two stages, we find that when N_0 increases from 1300 to 1700 with interval 100, their performances are very close, and the performance of $N_0 = 1300$ is marginally better than others. Also for the sake of computational efficiency, we select the first $N_0 = 1300$ best discriminant coordinates in our following experiments.

However, it should point out that the choice of N_0 is not necessarily easy and needs further research. The best N_0 may be not the same for different databases, and it may be ascertained by the cross validation method. We use the natural method and generalize the same $N_0 = 1300$ to other databases, in order to show that the LDC algorithm has robustness with respect to N_0. Figure 6 also validates the conclusion.

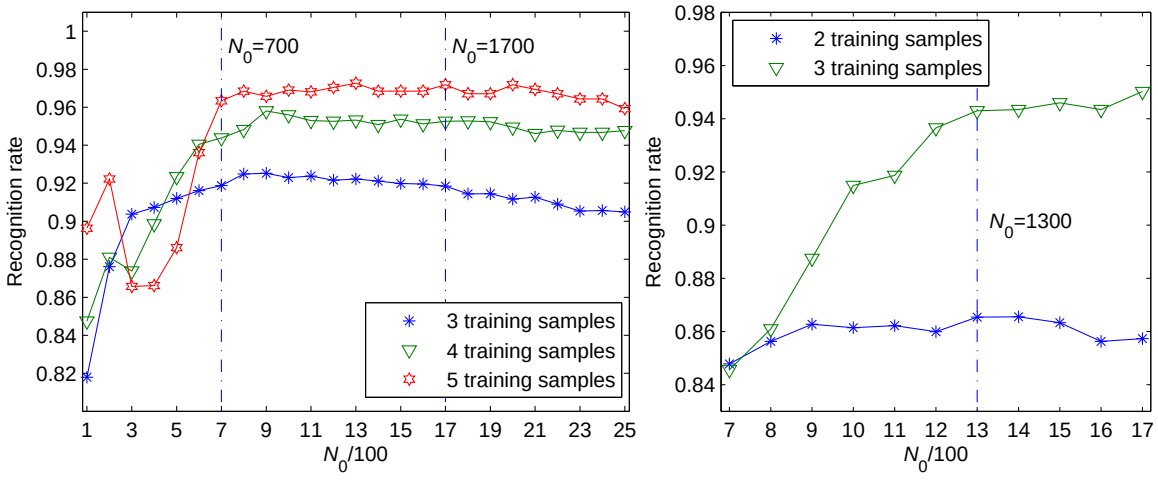

Figure 5: *(Left)* The performances of the LDC algorithm using different N_0 between 100 and 2500 with interval 100 on the small FERET data set. *(Right)* The performances of the LDC algorithm using different N_0 between 700 and 1700 with interval 100 on the large FERET data set.

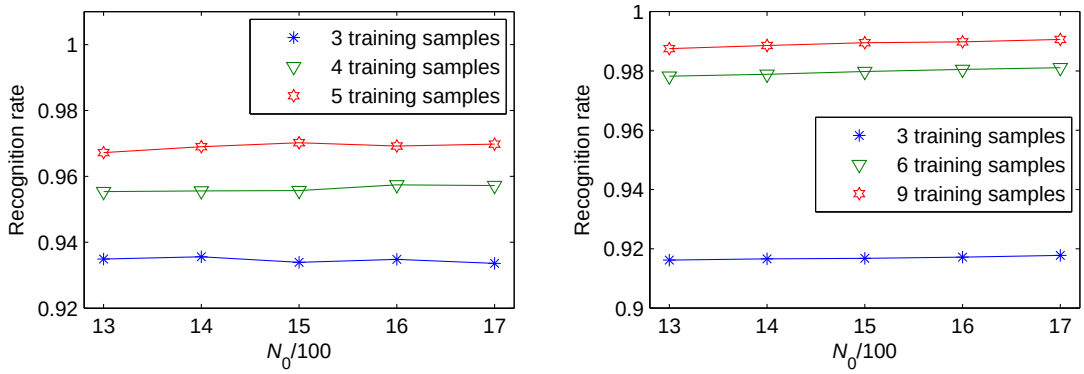

Figure 6: The performances of the LDC algorithm using different N_0 between 1300 and 1700 with interval 100. *(Left)* The ORL database. *(Right)* The CMU-lights database.

4.3.3 RECONFIRMATION OF THE DICTIONARY $\mathcal{D}$ USING N_0=1300

Moreover, we return to the anterior experiment (the construction of the dictionary $\mathcal{D}$) with the top $N_0(=1300)$ coordinates. The results prove that the third level has the best performance once again, as shown in Figure 7.

4.4 Efficacy of the LDC Based Feature Extraction

It is of paramount importance for face recognition to extract most discriminant features that are less sensitive to environmental variations. In this subsection, we compare the efficacy of feature

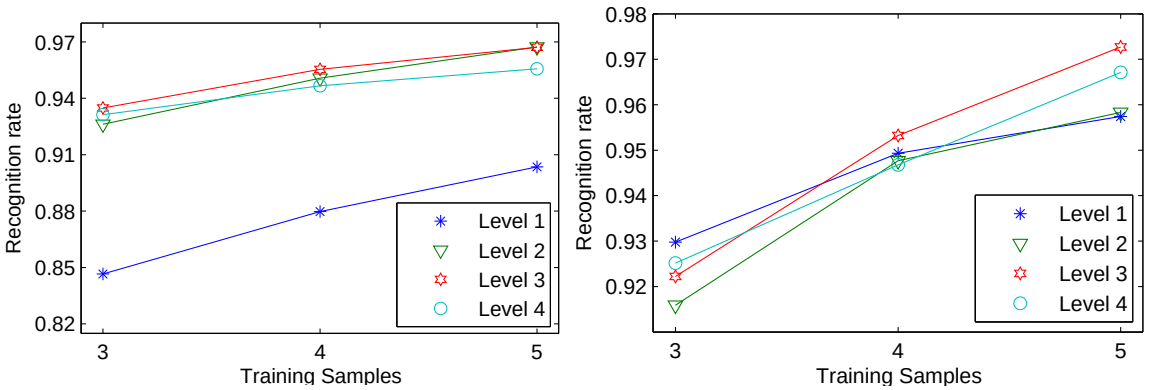

Figure 7: Performances of the LDC algorithm with N_0(=1300) coordinates in four levels. *(Left)* The ORL database. *(Right)* The small FERET data set.

Database	Training samples (i_0)	Methods with the l^2 criterion			
		LDC	LDB	MLDB	WaveletFace
ORL	3	93.43 ± 1.95	93.30 ± 2.22	92.49 ± 1.91	92.92± 1.60
	4	95.81 ± 1.59	95.60 ± 1.53	95.18 ± 1.39	94.56 ± 1.77
	5	96.95 ± 1.59	96.65 ± 1.20	96.43 ± 1.29	94.20 ± 1.82
FERET (small)	3	89.63 ± 3.01	87.56 ± 3.29	84.23 ± 3.26	88.80± 3.47
	4	92.92 ± 4.57	91.90 ± 4.48	88.80 ± 4.11	85.74 ± 4.85
	5	96.11 ± 3.39	95.69 ± 3.29	93.01 ± 3.80	87.50 ± 3.28
FERET (large)	2	79.53 ± 3.23	76.05 ± 2.90	67.16 ± 3.36	73.18± 4.77
	3	89.32 ± 1.82	87.45 ± 1.94	79.69 ± 1.99	72.59 ± 4.14
CMU-lights	3	79.46 ± 10.95	79.92 ± 10.91	81.30 ± 9.87	78.44± 10.92
	6	93.48 ± 7.63	93.69 ± 7.52	94.16 ± 6.89	85.08 ± 7.79
	9	96.83 ± 4.29	96.54 ± 4.50	96.76 ± 4.56	54.80 ± 17.52

Table 5: Comparison with wavelet-based methods

extraction in LDC with other wavelet-based methods, such as LDB, MLDB and WaveletFace on the ORL database, both data sets of the FERET database and the CMU-lights database. The setting of the feature extractions can be seen in Subsection 4.2. In order to show more comparability, *all the methods use CLDA and the NN classifier with the Euclidean distance*.

Table 5 shows that the LDC based feature extraction has the best result on the ORL database, both data sets of the FERET database, and it outperforms WaveletFace, though it underperforms LDB and MLDB marginally on the CMU-lights database. Moreover, LDC is more efficient than LDB, MLDB because of the lower computational complexity when K is large, especially on the large FERET data set ($K = 255$). On the whole, the feature extraction of LDC is more effective than the kin methods, including LDB, MLDB and WaveletFace.

Database	Training samples (i_0)	Methods with the *triangle square ratio* criterion			
		LDC	LDB	MLDB	WaveletFace
ORL	3	93.49 ± 2.19	92.93 ± 2.17	92.27 ± 1.82	92.80 ± 1.72
	4	95.54 ± 1.45	95.32 ± 1.51	94.67 ± 1.64	94.13 ± 1.82
	5	96.72 ± 1.71	96.65 ± 1.54	96.37 ± 1.41	93.60 ± 1.83
FERET (small)	3	92.22 ± 2.18	90.03 ± 2.39	87.15 ± 2.91	90.54 ± 2.29
	4	95.32 ± 1.90	94.31 ± 2.58	91.53 ± 3.31	87.78 ± 2.92
	5	97.27 ± 2.01	95.56 ± 2.56	94.17 ± 2.88	87.18 ± 2.21
FERET (large)	2	86.54 ± 2.78	84.11 ± 2.82	77.39 ± 3.34	79.01 ± 4.28
	3	94.30 ± 0.37	92.39 ± 0.48	87.87 ± 1.78	74.98 ± 5.17
CMU-lights	3	91.62 ± 6.63	91.90 ± 6.57	92.31 ± 6.37	88.67 ± 8.28
	6	97.82 ± 3.18	97.82 ± 3.25	98.14 ± 2.90	87.57 ± 6.86
	9	98.75 ± 2.09	98.47 ± 2.30	98.73 ± 2.14	55.96 ± 17.81

Table 6: Efficacy of the triangle square ratio criterion

4.5 Efficacy of the Triangle Square Ratio Criterion

Classifier and its classifying criterion are also important elements for face recognition. Generally, distance-based criterion is more robust than correlation-based criterion with respect to pose and expression changes while the contrary result is shown with respect to illumination variations. To extend the capacity covering variations of pose, expression and illumination, we have proposed the new triangle square ratio criterion in Subsection 3.5. In this experiment, *we use the same four feature extractions and CLDA as in Subsection 4.4, but the Euclidean distance is replaced by the triangle square ratio criterion for the NN classifer.* The results on the ORL database, both data sets of the FERET database and the CMU-lights database are shown in Table 6.

Comparing the results on Table 6 with Table 5 which uses the Euclidean distance, it shows that the triangle square ratio criterion performs better than the Euclidean distance considerably on both data sets of the FERET database and the CMU-lights database, while its efficacy is very close to the Euclidean distance on the ORL database. In fact, the FERET database, the CMU-lights database concern about illumination variations (light intensity and direction respectively), and the ORL database concerns about expression and pose changes. Comparison results show that the triangle square ratio criterion is more robust against illumination variations than the Euclidean distance, whilst retaining the robustness against pose and expression changes as the Euclidean distance.

Furthermore, *we replace the triangle square ratio criterion with the cosine criterion, whilst keeping the other setting,* on the ORL database and the small FERET data set. The performance of the cosine criterion is showed in Table 7. The comparison between Table 6 and Table 7 shows that the triangle square ratio criterion performs better than the cosine criterion considerably on the ORL database, although it marginally underperforms the cosine criterion on the small FERET data set.

4.6 Performance of the LDC Algorithm

In this part, we compare the performance of LDC with statistical methods, including FisherFace, PCA+CLDA, DLDA, and wavelet based methods: WaveletFace, LDB, MLDB on both individual and hybrid databases. *All the methods used for comparison keep their settings as their original forms (see Subsection 4.2).*

Database	Training samples (i_0)	Methods with the *cosine* criterion			
		LDC	LDB	MLDB	WaveletFace
ORL	3	92.54 ± 1.93	91.56 ± 2.12	90.99 ± 2.03	91.88 ± 1.73
	4	94.78 ± 1.42	94.25 ± 1.50	93.71 ± 1.71	93.46 ± 1.89
	5	96.22 ± 1.60	96.00 ± 1.49	95.53 ± 1.58	93.73 ± 1.76
FERET (small)	3	92.41 ± 1.97	90.15 ± 2.56	86.99 ± 3.00	90.82 ± 2.58
	4	95.67 ± 1.96	94.86 ± 2.31	91.85 ± 3.29	87.94 ± 3.51
	5	97.27 ± 2.01	95.56 ± 2.43	94.49 ± 2.62	87.55 ± 2.59

Table 7: Performance of the *cosine* criterion

Since our motivation is to compensate for illumination, pose and expression variation, from the properties of various databases, the ORL database is used to test moderate variations in pose and expression, the CMU-lights database to test illumination variations, the FERET database for more generic situation, and the hybrid database to test heteroscedastic class covariance distribution tolerance.

4.6.1 COMPARISON ON THE INDIVIDUAL DATABASES

The comparison of results are depicted in Table 8. Although some algorithms occasionally have better performance, LDC shows stable performance for every number of training samples per class on all databases. Especially on the large FERET data set, it outperforms FisherFace, PCA+CLDA, DLDA, WaveletFace, LDB, MLDB by 30.94%, 10.49%, 31.85%, 13.36%, 10.49%, 19.38% respectively when two samples per class are used for training, and by 21.75%, 5.95%, 31.75%, 21.71%, 6.85%, 14.61% respectively when three samples per class are used for training. In particular, LDC significantly outperforms FisherFace, DLDA, WaveletFace. From the standard derivation, we can see that LDC has better stability than other algorithms.

4.6.2 COMPARISON ON THE HYBRID DATABASE

On the hybrid database, we randomly select i_0 (i_0=2 to 4) images from each person for training, the rest ($6-i_0$) images of each individual in the small FERET data set are tested while the rest ($10-i_0$) images of each individual in the ORL database are used for testing. The comparison results are recorded in Table 9.

It shows that when the number of training sample per class increases from 2 to 4, the average recognition rates of LDC are from 82.20% to 90.28%. The performance is better than FisherFace, PCA+CLDA, DLDA, WaveletFace, LDB and MLDB which increase from 60.38%, 79.06%, 66.86%, 77.43%, 77.99%, 76.00%, to 75.63%, 90.64%, 80.85%, 66.09%, 88.11%, 86.09% respectively on the hybrid database.

As a whole, these experimental results reveal that LDC has better performance and stability on the ORL, FERET, CMU-lights databases than other methods, including FisherFace, PCA+CLDA, DLDA, WaveletFace, LDB and MLDB, especially outperforms FisherFace, DLDA, WaveletFace.

Database	Methods	Training samples (i_0)		
		3	4	5
ORL	FisherFace	87.20 ± 1.96	90.53 ± 1.87	92.07 ± 1.97
	PCA+CLDA	91.89 ± 1.97	94.81 ± 1.68	96.48 ± 1.58
	DLDA	84.17 ± 2.23	87.31 ± 1.92	90.17 ± 1.55
	WaveletFace	92.92 ± 1.60	94.56 ± 1.77	94.20 ± 1.82
	LDB	93.30 ± 2.22	95.60 ± 1.53	96.65 ± 1.20
	MLDB	92.49 ± 1.91	95.18 ± 1.39	96.43 ± 1.29
	LDC(TSR)	93.49 ± 2.19	95.54 ± 1.45	96.72 ± 1.71
FERET (small)	FisherFace	84.85 ± 3.64	88.01 ± 4.91	91.94 ± 4.23
	PCA+CLDA	89.07 ± 2.88	92.85 ± 4.06	95.60 ± 3.95
	DLDA	80.45 ± 4.81	86.34 ± 5.28	88.61 ± 6.41
	WaveletFace	88.80 ± 3.47	85.74 ± 4.85	87.50 ± 3.28
	LDB	87.56 ± 3.29	91.90 ± 4.48	95.69 ± 3.29
	MLDB	84.23 ± 3.26	88.80 ± 4.11	93.01 ± 3.80
	LDC(TSR)	92.22 ± 2.18	95.32 ± 1.90	97.27 ± 2.01
		Training samples (i_0)		
		2	3	
FERET (large)	FisherFace	55.60 ± 5.10	72.55 ± 2.57	
	PCA+CLDA	76.05 ± 3.36	88.35 ± 1.74	
	DLDA	54.69 ± 4.86	62.55 ± 2.30	
	WaveletFace	73.18 ± 4.77	72.59 ± 4.14	
	LDB	76.05 ± 2.90	87.45 ± 1.94	
	MLDB	67.16 ± 3.36	79.69 ± 1.99	
	LDC(TSR)	86.54 ± 2.78	94.30 ± 0.37	
		Training samples (i_0)		
		3	6	9
CMU-lights	FisherFace	80.57 ± 8.96	94.46 ± 6.44	97.36 ± 4.08
	PCA+CLDA	78.48 ± 10.89	93.91 ± 7.72	97.45 ± 3.70
	DLDA	76.92 ± 7.48	87.49 ± 6.60	92.65 ± 4.60
	WaveletFace	78.44 ± 10.92	85.08 ± 7.79	54.80 ± 17.52
	LDB	79.92 ± 10.91	93.69 ± 7.52	96.54 ± 4.50
	MLDB	81.30 ± 9.87	94.16 ± 6.89	96.76 ± 4.56
	LDC(TSR)	91.62 ± 6.63	97.82 ± 3.18	98.75 ± 2.09

Table 8: Comparison on the individual databases

4.7 Some Further Researches of the LDC Algorithm

In this subsection, we *keep the experimental settings in Subsection 4.2* and carry out some further researches of the LDC algorithm, including: effects of different wavelets, effects of different relative "distance" measures and comparison of CPU time.

Methods	Training samples (i_0)		
	2	3	4
FisherFace	60.38 ± 3.24	72.58 ± 2.51	75.63 ± 2.87
PCA+CLDA	79.06 ± 2.22	86.99 ± 1.34	90.64 ± 1.77
DLDA	66.86 ± 2.57	74.55 ± 2.61	80.85 ± 1.89
WaveletFace	77.43 ± 2.28	76.21 ± 1.55	66.09 ± 2.86
LDB	77.99 ± 2.21	85.81 ± 1.73	88.11 ± 1.78
MLDB	76.00 ± 2.07	83.41 ± 1.85	86.09 ± 2.20
LDC(TSR)	82.20 ± 1.69	88.46 ± 1.44	90.28 ± 1.47

Table 9: Comparison on the hybrid database

Database	wavelets	Training samples (i_0)		
		3	4	5
ORL	*harr*	92.49 ± 2.16	94.79 ± 1.55	96.33 ± 1.53
	db6	92.86 ± 1.52	95.24 ± 1.53	96.30 ± 1.55
	sym2	92.83 ± 2.02	95.29 ± 1.74	96.78 ± 1.77
	coif2	93.27 ± 1.37	94.94 ± 1.39	95.67 ± 1.67
	bior2.4	92.38 ± 2.08	94.04 ± 1.37	94.93 ± 1.54
	rbio2.4	93.71 ± 2.10	95.19 ± 1.33	95.78 ± 1.64
	db4	93.49 ± 2.19	95.54 ± 1.45	96.72 ± 1.71
FERET (small)	*harr*	93.69 ± 2.07	95.86 ± 2.17	97.22 ± 1.96
	db6	91.76 ± 2.65	94.58 ± 2.44	95.42 ± 2.42
	sym2	92.56 ± 2.41	95.72 ± 2.60	97.22 ± 2.34
	coif2	91.73 ± 2.55	94.75 ± 2.15	96.30 ± 1.94
	bior2.4	90.88 ± 2.10	94.47 ± 2.78	95.46 ± 1.86
	rbio2.4	91.59 ± 2.20	95.12 ± 2.37	96.16 ± 2.55
	db4	92.22 ± 2.18	95.32 ± 1.90	97.27 ± 2.01

Table 10: Performances of different wavelet basis functions

4.7.1 EFFECTS OF DIFFERENT WAVELET BASIS FUNCTIONS

We use different wavelet basis functions for the wavelet packet decomposition on the ORL database and the small FERET data set, including: *harr* wavelet, Daubechies *db6* wavelet, Symlets *sym2* wavelet, Coiflets *coif2* wavelet, Biorthogonal spline *bior2.4* wavelet, Reverse biorthogonal spline *rbio2.4* wavelet. Their performances are depicted in Table 10.

Table 10 shows that the performances of the *harr* and *sym2* wavelets are very close to the *db4* wavelet. Although other wavelets a little underperform the *db4* wavelet, their performances are also better than some other methods shown in Table 8. So we can conclude that the changes among aforementioned different wavelet basis functions have small effects on the performance of the LDC algorithm. Moreover, the orthogonal wavelets are superior to the biorthogonal wavelets in general.

4.7.2 EFFECTS OF DIFFERENT RELATIVE "DISTANCE" MEASURES

In order to show the effect of different relative "distance" measures on the performance of classification, we compare the dilation invariant entropy with the dilation invariant l^2 norm (15) on the

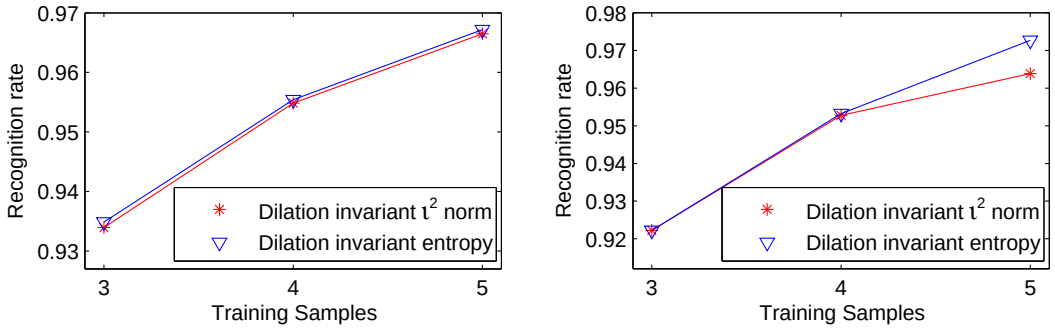

Figure 8: Performances of the LDC algorithm using the dilation invariant entropy and the dilation invariant l^2 norm. *(Left)* The ORL database. *(Right)* The small FERET data set.

Database	Methods				
	LDC	FisherFace	LDB	MLDB	WaveletFace
ORL ($K = 40$)	154	8	336	263	23
SmallFERET ($K = 72$)	164	24	415	400	40
LargeFERET ($K = 255$)	269	521	2888	3045	123

Table 11: Comparison of training CPU time (seconds)

ORL database and the small FERET data set. The results are shown in Figure 8. It shows that the dilation invariant l^2 norm marginally underperforms the dilation invariant entropy, which implies the changes between aforementioned different relative "distance" measures have slight effects on the performance of the LDC algorithm.

4.7.3 COMPARISON OF CPU TIME

We conduct an experiment to compare the time-consumption of the LDC algorithm with the popular statistics-based method: FisherFace and the wavelet-based methods: LDB, MLDB, WaveletFace on the ORL database and both data sets of the FERET database. We randomly select 3 images from each person for training. The experiments are implemented using MATLAB in a personal computer with Pentium 4 CPU and 256MB RAM. The time-consumptions are shown in Table 11. Although LDC is less efficient than WaveletFace, it is considerably more efficient than LDB and MLDB, especially when K is large. When K increases, the time-consumption of LDC increases more slowly than that of FisherFace, so that LDC can catch up with and surpass the efficiency of FisherFace.

It should point out that in our experiments, LDB selects the four best discriminant subbands from the local discriminant basis (see Subsection 4.2) and the number of selected coordinates is bigger than $N_0 (= 1300)$. So LDB takes more time than LDC, MLDB in the CLDA transform. However, MLDB based feature extraction needs to estimate the probability density functions, when K increases, it takes more and more time than LDB based feature extraction, so the total time in Table 11 shows that LDB is more expensive than MLDB on the ORL database and the small FERET data set due to their small K, the contrary result is shown on the large FERET data set due to its large K.

5. Discussions and Conclusion

In this paper, we have presented a novel local discriminant coordinates (LDC) method based on wavelet packet for face recognition to compensate for illumination, pose and expression variations. The method searches for the most discriminant coordinates from a wavelet packet dictionary, instead of the most discriminant basis as in the LDB algorithm. The LDC idea makes use of the scattered characteristic of the best discriminant features. In our method, the feature selection procedure is independent of subbands, and only depends on the discriminability of all coordinates. We have shown that the traditional "distance" measures (e.g., the l^2 distance, relative entropy) are deficient to measure the separability, while comparing the separability of two sample ensembles. We have proposed a new dilation invariant entropy which is independent of the order of magnitude. We have used the dilation invariant entropy and a MAP logistic model to measure the separability of coordinate-loading ensemble accurately. It locates in either low spatial frequency subbands or high spatial frequency subbands. So any two coordinates in the wavelet packet dictionary are comparable for their discriminability. Experimental results show that the LDC based feature extraction is more effective than LDB, MLDB, WaveletFace, PCA for feature extraction.

The LDC based feature extraction not only selects low frequency components, but also middle frequency components. From its significant improvement upon the WaveletFace method which only uses low frequency components, we can conclude that middle frequency components are helpful for face recognition, since their judicious combination with low spatial frequency components can improve the performance of face recognition greatly.

We have modified the Euclidean distance and the cosine criterion in the nearest neighbor classifier, and proposed a new triangle square ratio criterion which takes into account of two similarity measures, distance and correlation. Experimental results show that the triangle square ratio criterion is more robust against illumination variations than the Euclidean distance, while retaining the robustness against pose and expression changes as the Euclidean distance. Also, it can obviously outperform the cosine criterion when there are changes of pose and expression, although it marginally underperforms the cosine criterion when there are variations of illumination. So it can well extend the capacity covering variations of pose, expression and illumination.

We have used the ORL database to test moderate variations in pose and expression, the CMU database to test illumination variations, the FERET database for more generic situation, and the hybrid to test heteroscedastic class covariance distribution.

In conclusion, experimental results show that our LDC algorithm can well preserve the most discriminant information of facial image and improve the performance for face recognition under different variations. Also, experimental results show that our LDC algorithm has robustness with respect to the number of selected coordinates. The changes among some different wavelet basis functions and different relative "distance" measures have few effects on its performance. Moreover, it is an efficient method.

The LDC idea may have numerous applications beyond the one described in this paper. It also can be applied to feature or variable selection in other dictionaries of basis functions instead of wavelets, such as the local trigonometric functions.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable comments and suggestions, which improve the paper. This project is supported in part by NSF of China (60575004, 10231040), NSF of GuangDong(05101817), the Ministry of Education of China (NCET-04-0791) and the Hong Kong Research Grant Council (CityU 122506).

References

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces versus Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, Jul. 1997.

R. Bhagavatula and M. Savvides. PCA vs. automatically pruned wavelet-packet PCA for illumination tolerant face recognition. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 69–74, 2005.

L. F. Chen, H. Y. M. Liao, J. C. Lin, M. D. Kao, and G. J. Yu. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726, 2000.

J. T. Chien and C. C. Wu. Discriminant waveletfaces and nearest feature classifiers for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1644–1649, 2002.

W. Chou. Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition. In *Proceedings of IEEE*, volume 88, Aug. 2000.

R. R. Coifman and N. Saito. Constructions of local orthonormal bases for classification and regression. *Comptes Rendus Academy Science Paris*, 319:191–196, 1994.

R. R. Coifman and M. V. Wicherhauser. Entropy-based algorithm for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, Mar. 1992.

R. R. Coifman, Y. Meyer, and M. V. Wickerhauser. Wavelet analysis and signal processing. In M. B. Ruskai et al., editor, *Wavelets and Their Applications*, pages 153–178. Jones and Barlett, Boston, 1992.

D. Q. Dai and P. C. Yuen. Wavelet based discriminant analysis for face recognition. *Applied Mathematics and Computation*, 175:307–318, 2006.

D. Q. Dai and P. C. Yuen. Regularized discriminant analysis and its applications to face recognition. *Pattern Recognition*, 36(3):845–847, 2003.

I. Daubechies. The wavelet transform, time-frequency localization and signal processing. *IEEE Transactions on Information Theory*, 36:961–1005, 1990.

H. K. Ekenel and B. Sanker. Multiresolution face recognition. *Image and Vision Computing*, 23: 469–477, 2005.

G. C. Feng, P. C. Yuen, and D. Q. Dai. Human face recognition using PCA on wavelet subband. *Journal of Electronic Imaging*, 9(2):226–233, Apr. 2000.

K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, Boston, second edition, 1990.

L. Grewe and R. R. Brooks. On localization of objects in the wavelet domain. In *Proceedings of the 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 412–418, 1997.

A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

W. Jiang, G. H. Er, Q. H. Dai, and J. W. Gu. Similarity-based online feature selection in content-based image retrieval. *IEEE Transactions on Image Processing*, 15(3):702–712, 2006.

A. Z. Kouzani, F. He, and K. Sammut. Wavelet packet face representation and recognition. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 2, pages 1614–1619, Oct. 1997.

S. Kullback and R. A. Leibler. On information and sufficieny. *Annals of Mathematical Statistics*, 22:79–86, 1951.

M. Loog and R. P. W. Duin. Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):732–739, 2004.

S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

A. M. Martinez and M. Zhu. Where are linear feature extraction methods applicable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1934–1944, 2005.

S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K. R. Müller. Fisher discriminant analysis with kernels. In *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing IX*, pages 41–48, Aug. 1999.

K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

C. Nastar and N. Ayach. Frequency-based nonrigid motion analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:1067–1079, 1996.

J. Wang P. Howland and H. Park. Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition*, 39:277–287, 2006.

P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22: 1090–1104, 2000.

J. Weston B. Schölkopf A. Smola S. Mika, G. Rätsch and K. R. Müller. Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):623–628, 2003.

N. Saito and R. R. Coifman. Local discriminant bases. In *Proceedings of SPIE*, volume 2303, pages 2–14, 1994.

N. Saito and R. R. Coifman. Local discriminant bases and their applications. *Journal of Mathematical Imaging and Vision*, 5(4):337–358, 1995.

N. Saito, R. R. Coifman, F. B. Geshwind, and F. Warner. Discriminant feature extraction using empirical probability density estimation and a local basis library. *Pattern Recognition*, 35:2841–2852, 2002.

T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, Dec. 2003.

D. J. Stranss, G. Steidl, and W. Delb. Feature extraction by shape-adapted local discriminant bases. *Signal Processing*, 83:359–376, 2003.

P. P. Vaidyanathan. Multirate systems and filter banks. *Prentice-Hall, Englewood CliCs, NJ*, 1993.

N. Vaswani and R. Chellappa. Principal components null space analysis for image and video classification. *IEEE Transactions on Image Processing*, 15(7):1816–1830, 2006.

L. W. Wang, X. Wang, and J. F. Feng. Subspace distance analysis with application to adaptive bayesian algorithm for face recognition. *Pattern Recognition*, 39:456–464, 2006.

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, 2006.

J. Yang and J. Y. Yang. Why can LDA be performed in PCA transformed space? *Pattern Recognition*, 36(2):563–566, 2003.

J. Yang, A. F. Frangi, J. Y. Yang, and D. Zhang. KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):230–243, Feb. 2005.

M. H. Yang. Kernel Eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods. In *Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 215–220, May 2002.

H. Yu and J. Yang. A direct LDA algorithm for high-dimensional data–with application to face recognition. *Pattern Recognition*, 34(10):2067–2070, 2001.

B. L. Zhang, H. H. Zhang, and S. S. Ge. Face recognition by applying wavelet subband representation and kernel associative memory. *IEEE Transactions on Neural Networks*, 15(1):166–177, 2004.

J. Zhang and R. R. Korfhage. A distance and angle similarity measure method. *Journal of the American Society for Information Science*, 50(9):772–778, 1999.

Z. B. Zhang, S. L. Ma, and D. Y. Wu. The application of neural network and wavelet in human face illumination compensation. *Lecture Notes in Computer Science*, ISSU 3497, 2005.

W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–459, 2003.

X. S. Zhuang and D. Q. Dai. Improved discriminate analysis for high dimensional data and its application to face recognition. *Pattern Recognition*, 40(5):1570–1578, 2007.