Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 30274, Pages 1–11
DOI 10.1155/ASP/2006/30274

Information Theory for Gabor Feature Selection for Face Recognition

Linlin Shen and Li Bai

School of Computer Science and Information Technology, The University of Nottingham, Nottingham NG8 1BB, UK

Received 21 June 2005; Revised 23 September 2005; Accepted 26 September 2005

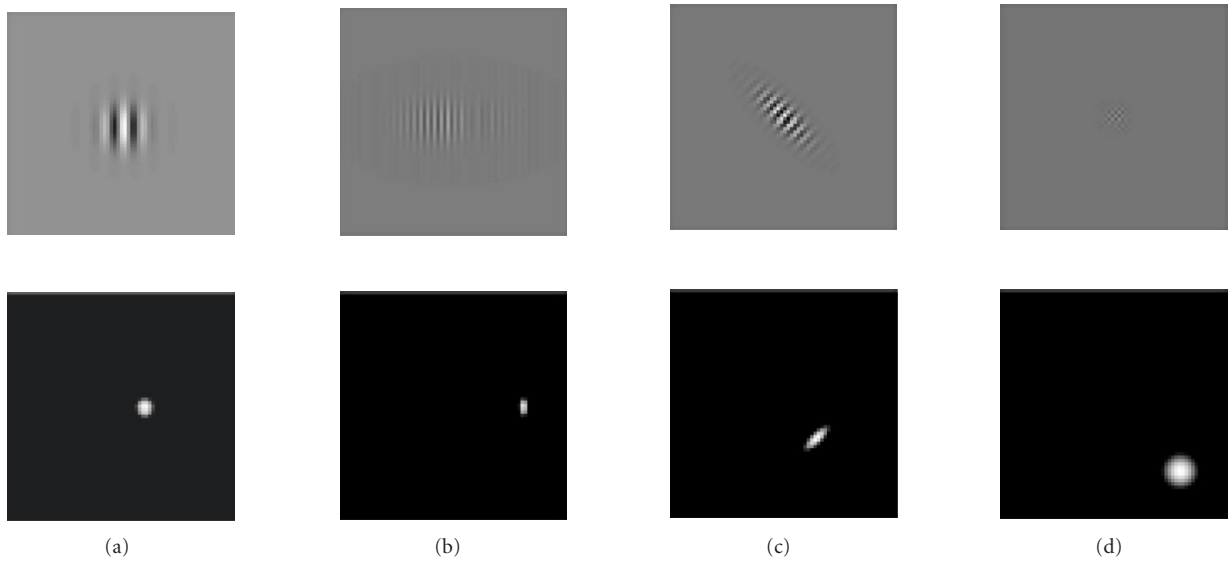
Recommended for Publication by Mark Liao

A discriminative and robust feature—kernel enhanced informative Gabor feature—is proposed in this paper for face recognition. Mutual information is applied to select a set of informative and nonredundant Gabor features, which are then further enhanced by kernel methods for recognition. Compared with one of the top performing methods in the 2004 Face Verification Competition (FVC2004), our methods demonstrate a clear advantage over existing methods in accuracy, computation efficiency, and memory cost. The proposed method has been fully tested on the FERET database using the FERET evaluation protocol. Significant improvements on three of the test data sets are observed. Compared with the classical Gabor wavelet-based approaches using a huge number of features, our method requires less than 4 milliseconds to retrieve a few hundreds of features. Due to the substantially reduced feature dimension, only 4 seconds are required to recognize 200 face images. The paper also unified different Gabor filter definitions and proposed a training sample generation algorithm to reduce the effects caused by unbalanced number of samples available in different classes.

1. INTRODUCTION

Daugman [1] presented evidence that visual neurons could optimize the general uncertainty relations for resolution in space, spatial frequency, and orientation. Gabor filters are believed to function similarly to the visual neurons of the human visual system. From an information-theoretic viewpoint, Okajima [2] derived Gabor functions as solutions for a certain mutual-information maximization problem. It shows that the Gabor receptive field can extract the maximum information from local image regions. Researchers have also shown that Gabor features, when appropriately designed, are invariant against translation, rotation, and scale [3]. Successful applications of Gabor filters in face recognition date back to the FERET evaluation competition [4], when the elastic bunch graph matching method [5] appeared as the winner. The more recent face verification competition [6] also saw the success of Gabor filters: both of the top two approaches used Gabor filters for feature extraction.

For face recognition applications, the number of Gabor filters used to convolve face images varies with applications, but usually 40 filters (5 scales and 8 orientations) are used [5, 7–9]. However, due to the large number of convolution operations of Gabor filters with the image (convolution at each position of the image), the computation cost is prohibitive. Even if a parallel system was used, it took about 7 seconds to convolve a 128×128 image with 40 Gabor filters [7]. For global methods (convolution with the whole image), the dimension of the feature vectors extracted is also incredibly large, for example, 163 840 for an image of size 64×64. To address this issue, a trial-and-error method is described in [10] that performs Gabor feature selection for facial landmark detection. A sampling method is proposed in [11] to determine the "optimal" position for extracting Gabor feature. This applies the same set of filters, which might not be optimal, at different locations of an image. Genetic algorithm (GA) has also been used to select Gabor features for pixel classification [12] and vehicle detection [13]. This basically creates a population of randomly selected combinations of features, each of which is considered a possible solution to the feature selection problem. However, the computation cost of GAs is very high, particularly in the case when a huge number of features are available. Recently, the AdaBoost algorithm has been used to select Haar-like features for face detection [14] and for learning the most discriminative Gabor features for classification [15]. Once the learning process is finished, Gabor filters of different frequencies and orientations are applied at different locations of the image for feature extraction.



(a) (b) (c) (d)

Figure 1: Gabor filters $\Pi(f, \theta, \gamma, \eta)$ in spatial domain (the 1st row) and frequency domain (the 2nd row), (a) $\Pi_a(0.1, 0, 1, 1)$; (b) $\Pi_b(0.3, 0, 6, 3)$; (c) $\Pi_c(0.2, \pi/4, 3, 1)$; (d) $\Pi_d(0.4, \pi/4, 2, 2)$.

Despite its success, AdaBoost algorithm selects only features that perform "individually" best, the redundancy among selected features is not considered [16]. In this paper, we present a conditional mutual-information-[17, 18] based method for selecting Gabor features for face recognition. A small subset of Gabor features capable of discriminating intrapersonal and interpersonal spaces is selected using the information theory, which is then subjected to generalized discriminant analysis (GDA) for class separability enhancement. The experimental results show that 200 features are enough to achieve highly competitive accuracy for the face database used. Significant computation and memory efficiency have been achieved since the dimension of features has been reduced from 163 840 to 200 for 64×64 images. The kernel enhanced informative Gabor features have also been tested on the whole FERET database following the same evaluation protocol and improved performance on three test sets has been achieved.

2. GABOR FEATURE EXTRACTION

2.1. Gabor filters

In the spacial domain, the 2D Gabor filter is a Gaussian kernel modulated by a sinusoidal plane wave [3]:

$$\begin{aligned} \varphi_{\Pi(f,\theta,\gamma,\eta)}(x, y) &= \frac{f^2}{\pi\gamma\eta} e^{-(\alpha^2 x'^2 + \beta^2 y'^2)} e^{j2\pi f x'}, \\ x' &= x\cos\theta + y\sin\theta, \\ y' &= -x\sin\theta + y\cos\theta, \end{aligned} \tag{1}$$

where f (cycles/pixel) is the central frequency of the sinusoidal plane wave, θ is the anticlockwise rotation of the Gaussian and the plane wave, α is the sharpness of the Gaussian along the major axis parallel to the wave, and β is the sharpness of the Gaussian minor axis perpendicular to the wave. $\gamma = f/\alpha$ and $\eta = f/\beta$ are defined such that the ratio between frequency and sharpness is constant. Figure 1 shows four Gabor filters with different parameters in both spatial domain and frequency domain.

Note that (1) is different from the one normally used for face recognition [5, 7–9], however, this equation is more general. Given that the orientation θ of the major axis of the elliptical Gaussian is the same as that of the sinusoidal plane wave, the wave vector k (radian/pixel) can now be expressed as $\vec{k} = 2\pi f \exp(j\theta)$. Setting $\gamma = \eta = \sigma/\sqrt{2}\pi$, that is, $\alpha = \beta = \sqrt{2}\pi f/\sigma$, the Gabor filter located at position $\vec{z} = (x, y)$ can now be defined as

$$\varphi(\vec{z}) = \frac{1}{2\pi} \frac{||\vec{k}||^2}{\sigma^2} \exp\left(\frac{-||\vec{k}||^2 ||\vec{z}||^2}{2\sigma^2}\right) \exp(i\vec{k}\cdot\vec{z}). \tag{2}$$

The Gabor functions used in [5, 7–9] have been derived from (1), which can be seen as a special case when $\alpha = \beta$. Similarly, the relationship between (1) and those in [10, 19] could also be established. When DC term could be deduced to make the wavelet DC free [5, 7–9], similar effects can also be achieved by normalizing the image to be zero mean [20].

2.2. Gabor feature representation

Once Gabor filters have been designed, image features at different locations, frequencies, and orientations can be extracted by convolving the image $I(x, y)$ with the filters:

$$O_{\Pi(f,\theta,\gamma,\eta)}(x, y) = I(x, y) * \varphi_{\Pi(f,\theta,\gamma,\eta)}(x, y). \tag{3}$$

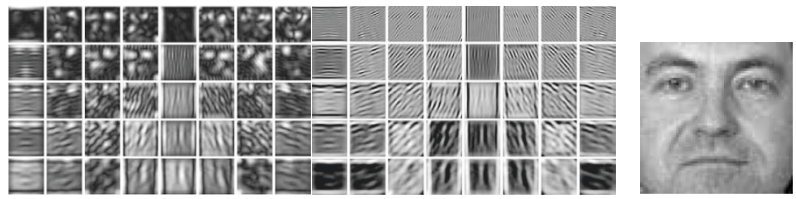


Figure 2: Magnitude and real part of an image convolved with 40 Gabor filters.

A number of Gabor filters at different scales and orientations are usually used. We designed a filter bank with 5 scales and 8 orientations for feature extraction [7]:

$$\{\varphi_{\Pi(f_u,\theta_v,\gamma,\eta)}(x,y)\}, \quad \gamma=\eta=0.8,\ f_u=\frac{f_{\max}}{\sqrt{2}^u}, \quad \theta_v=\frac{v}{8}\pi,\ u=0,\ldots,4,\ v=0,\ldots,7, \tag{4}$$

where f_u and θ_v define the orientation and scale of the Gabor filter, $f_{\max}$ is the maximum frequency, and $\sqrt{2}$ (half octave) is the spacing factor between different central frequencies. According to the Nyquist sampling theory, a signal containing frequencies higher than half of the sampling frequency cannot be reconstructed completely. Therefore, the upper limit frequency for a 2D image is 0.5 cycles/pixel, whilst the low limit is 0. As a result, we set $f_{\max}=0.5$. The resultant Gabor feature set thus consists of the convolution results of an input image $I(x,y)$ with all of the 40 Gabor filters:

$$S=\{O_{u,v}(x,y): u\in\{0,\ldots,4\},\ v\in\{0,\ldots,7\}\}, \tag{5}$$

where $O_{u,v}(x,y)=|I(x,y)*\varphi_{\Pi(f_u,\theta_v,\gamma,\eta)}(x,y)|$. Figure 2 shows the magnitudes of Gabor representation of a face image with 5 scales and 8 orientations. A series of row vectors $\mathbf{O}^I_{u,v}$ could be obtained out of $O_{u,v}(x,y)$ by concatenating its rows or columns, which are then concatenated to generate a discriminative Gabor feature vector:

$$G(I)=\mathbf{O}(I)=(\mathbf{O}^I_{0,0}\ \ \mathbf{O}^I_{0,1}\ \ \cdots\ \ \mathbf{O}^I_{4,7}). \tag{6}$$

Take an image of size 64×64 for example, the convolution result will give $64\times64\times5\times8=163\,840$ features. Each Gabor feature is thus extracted by a filter with parameters f_u, θ_v at location (x,y). Since the parameters of Gabor filters are chosen empirically, we believe a lot of redundant information is included, and therefore a feature selection mechanism should be used to choose the most useful features for classification.

3. MUTUAL INFORMATION FOR FEATURE SELECTION

3.1. Entropy and mutual information

As a basic concept in information theory, entropy $H(X)$ is used to measure the uncertainty of a random variable (rv) X. If X is a discrete rv, $H(X)$ can be defined as below:

$$H(X)=-\sum_x p(X=x)\lg\left(p(X=x)\right). \tag{7}$$

Mutual information $I(Y;X)$ is a measure of general interdependence between two random variables X and Y

$$I(Y;X)=H(X)+H(Y)-H(X,Y). \tag{8}$$

Using Bayes rule on conditional probabilities, (8) can be rewritten as

$$I(Y;X)=H(X)-H(X\mid Y)=H(Y)-H(Y\mid X). \tag{9}$$

Since $H(Y)$ measures the a priori uncertainty of Y and $H(Y\mid X)$ measures the conditional a posteriori uncertainty of Y after X has been observed, the mutual information $I(Y;X)$ measures how much the uncertainty of Y is reduced if X has been observed. It can be easily shown that if X and Y are independent, $H(X,Y)=H(X)+H(Y)$, and consequently their mutual information is zero.

3.2. Conditional mutual information

In the context of information theory, the aim of feature selection is to select a small subset of features $(X_{v(1)},X_{v(2)},\ldots,X_{v(K)})$ from $(X_1,X_2,\ldots,X_N)$ that gives as much information as possible about Y, that is, maximize $I(Y;X_{v(1)},X_{v(2)},\ldots,X_{v(K)})$. However, the estimation of this expression is unpractical since the number of probabilities to be decided could be as huge as 2^{K+1} even when the value of r.v. is binary. To address this issue, one approach is to use conditional mutual information (CMI) for feature fitness measurement. Given a set of candidate features $(X_1,X_2,\ldots,X_N)$, CMI $I(Y;X_n\mid X_{v(k)})$, $1\le n\le N$, could be used to measure the information about Y carried by the feature X_n when a feature $X_{v(k)}$, $k=1,2,\ldots,K$, is already selected:

$$\begin{aligned} I(Y;X_n\mid X_{v(k)}) &= H(Y\mid X_{v(k)})-H(Y\mid X_n,X_{v(k)}) \\ &= H(Y,X_{v(k)})-H(X_{v(k)}) \\ &\quad -H(Y,X_n,X_{v(k)})+H(X_n,X_{v(k)}). \end{aligned} \tag{10}$$

We can justify the fitness of a candidate feature by its CMI given an already selected feature, that is, a candidate feature is good only if it caries information about Y, and if this information has not been caught by any of the $X_{v(k)}$ already selected. When there are more than two selected features, the minimum CMI given each selected feature, that is, $\min_k I(Y;X_n\mid X_{v(k)})$, could be used as the fitness function.

For $j = 1, 2, \ldots m$
  For $u = 0, 1, \ldots 4$
    For $v = 0, 1, \ldots 7$
      Randomly generate an image pair (I_p, I_q) from different person
      Calculate the Gabor feature difference $\mathbf{Z}_{u,v}$ corresponding to filter $\varphi_{u,v}(x, y)$ using the image pair as below:
        $\mathbf{Z}_{u,v} = |\mathbf{O}_{u,v}^{I_p} - \mathbf{O}_{u,v}^{I_q}|$
    End
  End
  Concatenate the 40 feature differences into an extrapersonal sample,
  $g_j = [\, \mathbf{Z}_{0,0}\mathbf{Z}_{0,1} \cdots \mathbf{Z}_{u,v} \cdots \mathbf{Z}_{4,7} \,]$
End
Output the m extrapersonal Gabor feature difference samples
$\{(g_1, y_1), \ldots, (g_m, y_m)\}$, $y_1 = y_2 = \cdots = y_m = 1$

Algorithm 1: Extrapersonal training samples generation.

This selection process thus takes both individual strength and redundancy among selected features into consideration. The estimation of CMI requires information about the marginal distributions $p(X_n)$,$p(Y)$ and the joint probability distributions $p(Y, X_{v(k)})$, $p(X_n, X_{v(k)})$, and $p(Y, X_n, X_{v(k)})$, which could be estimated using a histogram. However, it is very difficult to determine the number of histogram bins. Though Gaussian distribution could be applied as well, many of the features, as shown in the experimental section, do not show the Gaussian property. To reduce the complexity and computation cost of the feature selection process, we hereby focus on random variables with binary values only, that is, $x_n \in \{0, 1\}$, $y \in \{0, 1\}$, where x_n and y are the values of random variables X_n and Y, respectively. For binary rv, the probability could be estimated by simply counting the number of possible cases and dividing that number with the total number of training samples. For example, the possible cases will be $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ for the joint probability of two binary random variables $p(Y, X_{v(k)})$.

4. SELECTING INFORMATIVE GABOR FEATURES

4.1. The Gabor feature difference space

Due to the complexity of estimation of CMI, the work presented here focuses on two-class problem only. As a result, the face recognition problem is formulated as a problem in the difference space [21] for feature selection, which models dissimilarities between two facial images. Two classes, dissimilarities between faces of the same person (intrapersonal space) and dissimilarities between faces of the different people (extrapersonal space), are defined. The two Gabor feature difference sets CI (intrapersonal difference) and CE (extrapersonal difference) can be defined as

$$CI = \{||G(I_p) - G(I_q)||,\ p = q\},$$
$$CE = \{||G(I_p) - G(I_q)||,\ p \neq q\}, \tag{11}$$

where I_p and I_q are the facial images from people p and q, respectively, and $G(\cdot)$ is the Gabor feature extraction operation as defined in last section. Each of the M samples in the difference space can now be described as $g_i = [x_1 x_2 \ \cdots \ x_n \ \cdots \ x_N]$, $i = 1, 2, \ldots, M$, where N is the dimension of extracted Gabor features and $x_n = (\|G(I_p) - G(I_q)\|)_n = (\|\mathbf{O}(I_p) - \mathbf{O}(I_q)\|)_n$.

4.2. Training samples generation

For a training set with L facial images captured for each of the D persons, $D(\frac{L}{2})$ samples could be generated for intrapersonal difference class while $(\frac{DL}{2}) - D(\frac{L}{2})$ samples are available for extrapersonal difference class. There are always much more extrapersonal samples than intrapersonal samples for face recognition problems. Take a database with 400 images from 200 subjects for example, 200 intrapersonal image pairs and $(\frac{400}{2}) - 200 = 79\,800$ extrapersonal image pairs are available. To achieve a balance between the numbers of training samples from the two classes, a random subset of the extrapersonal samples could be produced. However, we also want to make the subset a representative of the whole set as much as possible. To achieve this tradeoff, we proposed a procedure shown in Algorithm 1 to generate m extrapersonal samples using 40 (5 scales, 8 orientations) Gabor filters: instead of using only m pairs, our method randomly generates m samples from $m \times 40$ extrapersonal image pairs. As a result, without increasing the number of extrapersonal samples to bias the feature selection process, the training samples thus generated are more representative.

With $l = D(\frac{L}{2})$ intrapersonal difference samples, the training sample generation process finally outputs a set of $M = m + l$ Gabor feature difference samples: $\{(g_1, y_1), \ldots, (g_M, y_M)\}$. Each sample $g_i = [x_1 x_2 \ \cdots \ x_n \ \cdots \ x_N]$ in the difference space is associated with a binary label: $y_i = 0$ for an intrapersonal difference, while $y_i = 1$ for an extrapersonal difference.

4.3. Gabor feature selection using CMI

Once a set of training face samples with class label (intrapersonal, or extrapersonal) $\{(g_1, y_1), (g_2, y_2), \ldots (g_M, y_M)\}$, $g_i = [x_1 x_2 \ \cdots \ x_n \ \cdots \ x_N]$, is given, each feature of the sample in the difference space is now also converted to binary value as below, that is, if the difference is less than a threshold, the difference is set as 0, otherwise it is set as 1:

$$x_n = \begin{cases} 0, & x_n < t_n, \\ 1, & x_n \geq t_n. \end{cases} \tag{12}$$

```
Given a set of candidate features (X_1, X_2, ..., X_N)
 and sample labels Y
K = 1
v(K) = arg max_n I(Y; X_n)
while K < K_max
      for each candidate feature X_n
            calculate CMI I(Y; X_n | X_v(k)) given
             each of the selected feature
            X_v(k), k = 1, 2, ... K
      end
      v(K + 1) = arg max_n {min_k I(Y; X_n | X_v(k))}
      K = K + 1
end
```

Algorithm 2: CMI for feature selection.

Since we are only interested in the selection of features, the threshold t_n is simply determined by the centre of intrapersonal samples mean and extrapersonal samples mean:

$$t_n = \frac{1}{2}\left(\frac{1}{m}\sum_{p=1}^{m}\left((g_p)_n \mid y_p = 1\right) + \frac{1}{l}\sum_{q=1}^{l}\left((g_q)_n \mid y_p = 0\right)\right), \tag{13}$$

where m and l are the numbers of intra- and extrapersonal difference samples, respectively. Once the features are binarized, the set of training samples can now be represented by N binary random variables $(X_1, X_2, \ldots, X_N)$ representing candidate features and a binary random variable Y representing class labels. The iterative process listed in Algorithm 2 can be used to select the informative Gabor features. The Gabor features thus selected carry important information about predicting whether the sample is an intrapersonal difference or an extrapersonal difference. Based on the fact that face recognition is actually to find the most similar match with the least difference, the selected features will also be very important for recognition.

5. KERNEL ENHANCEMENT FOR RECOGNITION

Once the most informative Gabor features are selected, different approaches could be used for face recognition, for example, principal component analysis (PCA) or linear discriminant analysis (LDA) can be further applied for enhancement and the nearest-neighbor (NN) classifier can be used for classification. Recently, kernel methods have been successfully applied to solve pattern recognition problems because of their capacity in handling nonlinear data. By mapping sample data to a higher-dimensional feature space, effectively a nonlinear problem defined in the original image space is turned into a linear problem in the feature space [22]. Support vector machine (SVM) is a successful example of using the kernel methods for classification. However, SVM is basically designed for two-class problem and it has been shown in [23] that nonlinear kernel subspace methods perform better than SVM for face recognition. As a result, we use generalized discrimniant analysis (GDA) [24] for further feature enhancement and KNN classifier for recognition. GDA subspace is firstly constructed from the training image set and each image in the gallery set is projected onto the subspace. To classify an input image, the selected Gabor features are extracted and then projected to the GDA subspace. The similarity between any two facial images can then be determined by distance of the projected vectors. Different distance measures such as Euclidean, Mahalanobis, and normalized correlation have been tested in [9] and the results show that the normalized correlation distance measure is the most appropriate one for GDA method.

As a generalization of LDA, GDA performs LDA on sample data in the high-dimension feature space F via a nonlinear mapping ϕ. To make the algorithm computable in the feature space F, kernel method is adopted in GDA. Given that the dot product of two samples in the feature space can be easily computed via a kernel function, the computation of an algorithm in F can now be greatly reduced. By integrating the kernel function into the within-class variance S_w and between-class variance S_b of the samples in F, GDA can successfully determine the subspace to maximize the ratio between S_b and S_w. While the maximal dimension of LDA is determined by the number of classes C [25], the maximal dimension of GDA subspace is also determined by the rank of the kernel matrix K, that is, $\min\{C - 1, \text{rank}(K)\}$ [24].

6. EXPERIMENTAL RESULTS

We first analyze the performance of our algorithm using a subset of FERET database, which is a standard testbed for face recognition technologies [4]. Six hundred frontal face images corresponding to 200 subjects are extracted from the database for the experiments—each subject has three images of size 256×384 with 256 gray levels. The images were captured at different photo sessions so that they display different illumination and facial expressions. Two images of each subject are randomly chosen for training, and the remaining one is used for testing. Figure 3 shows the sample images from the database. The first two rows are the example training images while the third row shows the example test images.

The following procedures were applied to normalize the face images prior to the experiments.

(i) The centres of the eyes of each image are manually marked.
(ii) Each image is rotated and scaled to align the centres of the eyes.
(iii) Each face image is cropped to the size of 64×64 to extract facial region.
(iv) Each cropped face image is normalized to zero mean and unit variance.

Figure 3: Sample images used in experiments.

6.1. Selected Gabor features

The randomly selected 400 face images (2 images each subject) are used to learn the most important Gabor feature for intrapersonal and extrapersonal face space discriminations. As a result, 200 intrapersonal face difference samples and 1 600 extrapersonal face difference samples using the method as described in Section 4.2 are randomly generated for feature selection. When implemented in Matlab 6.1 and a P4 1.8 GHz PC, it took about 12 hours to select 200 features from the set of training data. Figure 4 shows the first six selected Gabor features and locations of the 200 Gabor features on a typical face image in the database. It is interesting to see that most of the selected Gabor features are located around the prominent facial features such as eyebrows, eyes, noses, and chins, which indicates that these regions are more robust against the variance of expression and illumination. This result is agreeable with the fact that the eye and eyebrow regions remain relatively stable when the person's expression changes. Figure 5 shows the distribution of selected filters in different scales and orientations. As shown in the figure, filters centred at low-frequency band are selected much more frequently than those at high-frequency band. On the other hand, majority of the discriminative Gabor features are with orientation around $3\pi/8$, $\pi/2$, and $5\pi/8$. The orientation preference indicates that horizontal features seem to be more important for face recognition task.

To check whether the distribution of the Gabor features in the difference space is Gaussian or not, we list in Table 1 the normalized skewness and kurtosis for each of the first 10 selected features. The hypothesis for the test is that a set of observations follows the Gaussian distribution if the normalized skewness and kurtosis of the data follow the standard Gaussian distribution $N(0,1)$ [26], which can be defined as below:

$$S = \frac{1}{\sqrt{6N}\sigma^3}\sum_{i=1}^{N}(x_i-\bar{x})^3,$$
$$K = \frac{1}{\sqrt{24N}\sigma^4}\sum_{i=1}^{N}(x_i-\bar{x})^4 - \sqrt{\frac{3N}{8}}, \tag{14}$$

where N, $\bar{x}$, σ are the sample size, sample mean, and sample standard deviation, respectively. Given the critical values for the standard Gaussian distribution as ± 1.96, we observe from Table 1 that all of the 10 features are non-Gaussian since their kurtosis exceeds the critical value. The information gain of the first 10 features has also been included in Table 1, for example, the value for the second feature shows the information carried by it when the first feature has been selected. As shown, the gain decreases monotonically when more features are included.

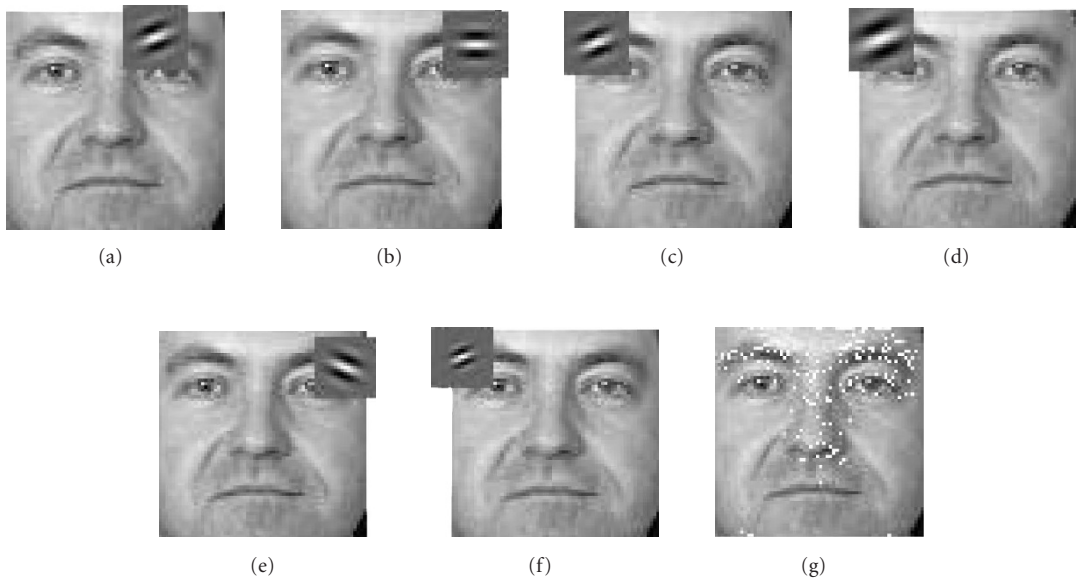


Figure 4: First six selected Gabor features (a)–(f); and the 200 selected feature points (g).

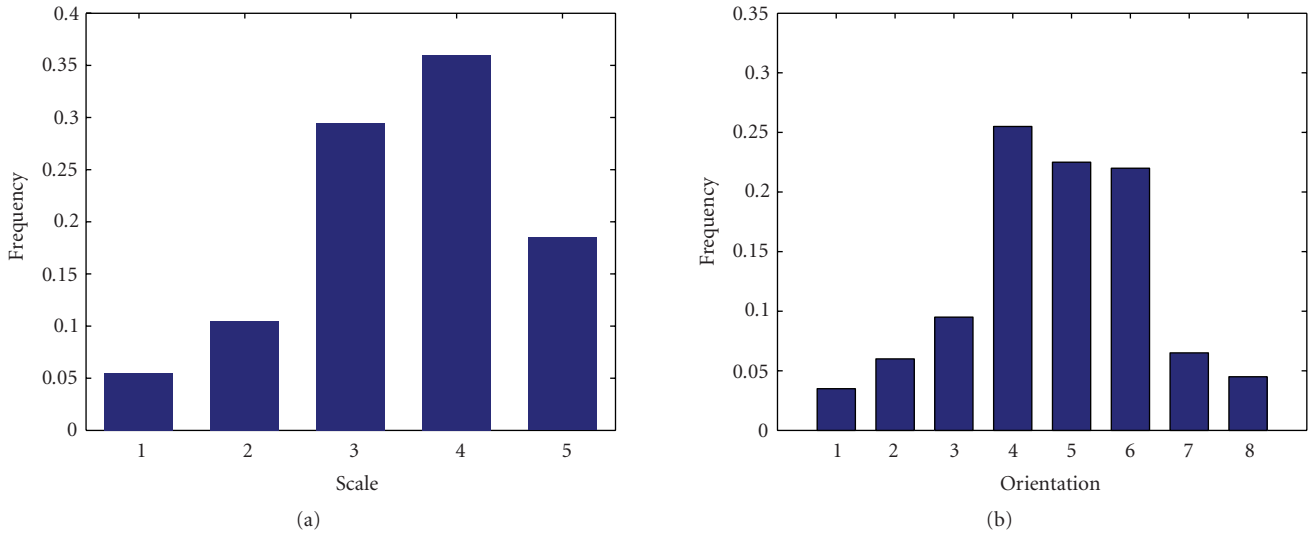


Figure 5: Distribution of selected filters in scale and orientation.

Table 1: Information gain, skewness, and kurtosis of the first 10 selected features.

Feature number	1	2	3	4	5	6	7	8	9	10
Information gain	0.1603	0.1253	0.1155	0.1084	0.1076	0.1017	0.1017	0.1009	0.0995	0.0994
Skewness	1.0548	1.2035	1.1914	1.0275	0.9540	1.0968	0.9865	1.0047	1.2664	1.1999
Kurtosis	3.6319	4.3834	4.2048	3.6621	3.5001	3.8315	3.4612	3.5050	4.2637	4.2075

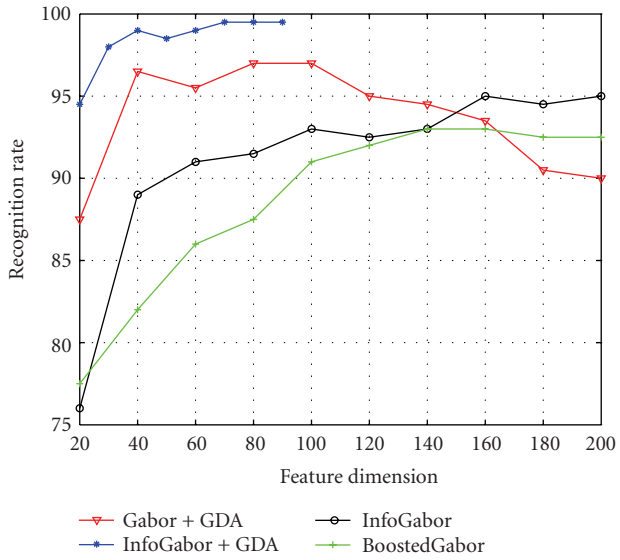


Figure 6: Recognition performance using different Gabor features.

6.2. Recognition performance on the subset of FERET database

Once the informative Gabor features (InfoGabor) are selected, we are now able to apply them directly for face recognition. Normalized correlation distance measure and 1-NN classifier are used. For comparison, we have also implemented the AdaBoost algorithm to select Gabor features for face recognition (BoostedGabor), using exactly the same training set. During boosting, exhaustive search is performed in the Gabor feature difference space as defined in (12). By picking up at each iteration the feature with the lowest weighted classification error, AdaBoost algorithm selects one by one those features that are significant for classification. As mentioned before, the features selected by AdaBoost perform "individually" well, but there are still lots of redundancy available. As a result, many features selected by AdaBoost are similar. Details of the learning process can be found in [15]. The performance shown in Figure 6 proves the advantage of InfoGabor over BoostedGabor. As shown in the figure, InfoGabor achieved as high as 95% recognition rate with 200 features. The performance drop using 120 features could be caused by the variance between test images and training images—some features significant to discriminate training images might not be the appropriate ones for test images. A more representative training set could alleviate this problem.

In the next series of experiments, we perform GDA on the selected Gabor features (InfoGabor-GDA) for face recognition. To show the robustness and efficiency of the proposed methods, we also perform GDA on the whole Gabor feature set (Gabor-GDA) for comparison purposes. Downsampling is adopted to reduce feature dimension to a certain level, see [9] for details. Normalized correlation distance measure and the nearest-neighbor classifier are used for both methods. The maximum dimensions of GDA subspace for InfoGabor-GDA and Gabor-GDA are 96 and 199, respectively. It can be observed from Figure 6 that InfoGabor-GDA performs a little better than Gabor-GDA. Accuracy of 99.5% is achieved when dimension of GDA space is set as 70, while Gabor-GDA needs 80 to achieve 97% accuracy. The comparison shows that some important Gabor features may have been missing during the dowsampling process, while many features that remained are, on the other hand, redundant. We also compare the computation and memory cost of Gabor-GDA and InfoGabor-GDA in Table 2. This shows that InfoGabor-GDA requires significantly less computation and memory than Gabor-GDA, for example, the number of convolutions to extract Gabor features is reduced from 16 3840 to 200. Although fast Fourier transform (FFT) could be used here to circumvent the convolution process, the feature extraction process still takes about 1.5 seconds in our C implementation whilst the 200 convolutions takes less than 4 milliseconds. For Gabor-GDA with downsample rate = 16, the feature dimension is reduced to 10 240, which is still 50 times of the dimension of InfoGabor-GDA. As a result, InfoGabor-GDA is much faster in training and testing. While it takes Gabor-GDA 275 seconds to construct the GDA subspace using the 400 training images, it takes InfoGabor-GDA only about 6 seconds. InfoGabor-GDA also achieves substantial recognition efficiency—only 4 seconds are required to recognize the 200 test images. The computation time is recorded in Matlab 6.1, with a P4 1.8 GHz PC.

Having shown in our previous work [9] that GDA achieved significantly better performance on the whole Gabor feature set (Gabor-GDA) than LDA (Gabor-LDA), we also performed LDA on the selected informative Gabor features (InfoGabor-LDA) for comparison. The results are shown in Figure 7, together with that of InfoGabor as a baseline. The results show that instead of enhancing it, the application of LDA surprisingly deteriorates the performance of InfoGabor. Only 80% accuracy is achieved when the dimension of LDA subspace is set as 60. The result suggests that when the input features are discriminative enough, LDA analysis may not necessarily lead to a more discriminative space. The results also show that the feature enhancement ability of GDA is better than LDA.

6.3. Recognition performance on the whole FERET database

We now test our InfoGabor-GDA algorithm on the whole FERET database. According to the FERET evaluation protocol, a gallery of 1196 frontal face images and 4 different prob sets are used for testing. The numbers of images in different prob sets are listed at Table 3, with example images shown in Figure 8. Fb and Fc prob sets are used for assessing the effect of facial expression and illumination changes, respectively, and there is only a few seconds between the capture of the gallery-probe pairs. Dup I and Dup II consist of images

Table 2: Comparative computation and memory cost of Gabor-GDA and InfoGabor-GDA.

Methods	Number of convolutions to extract Gabor feature	Dimension of Gabor features before GDA	Training time (s)	Test time (s)
Gabor-GDA	$64 \times 64 \times 40 = 163\,840$	10 240	275	263
InfoGabor-GDA	200	200	6	4

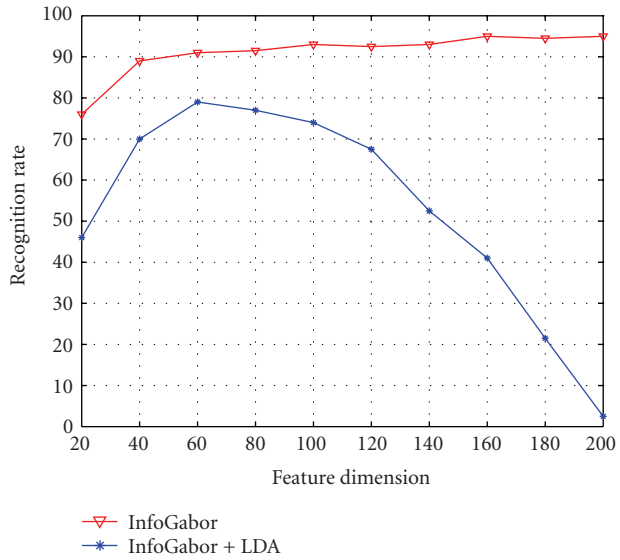


Figure 7: Recognition performance of InfoGabor-LDA.

taken on different days from their gallery images, and particularly, there is at least one year between the acquisition of the probe image in Dup II and the corresponding gallery image. A training set consisting of 736 images is used to select the most informative Gabor features and construct the GDA subspace [28]. As a result, 592 intrapersonal and 2000 extrapersonal samples are produced to select 300 Gabor features using the sample generation algorithm and information theory. The feature selection process took about 18 hours in Matlab 6.1, with a P4 1.8 GHz PC. During development phase, the training set is randomly divided into a gallery set with 372 images and a test set with 364 images to decide the RBF kernel and dimension of GDA for optimal performance. The same parameters are used throughout the testing process.

Performance of the proposed algorithm is shown in Table 4, together with that of the main approaches used in FERET evaluation [4], and the approach that extracts Gabor features from variable feature points [27]. The results show that our method achieves the best result on sets Fb, Fc, and Dup II due to the robustness of selected Gabor features against variation of expression, illumination, and capture time. Particularly, the performance of our methods is significantly better than all of other methods on Dup II. The elastic graph matching (EGM) method, based on the dynamic link architecture, performs a little better than our method on Dup I. However, the method requires intensive computation for both Gabor feature extraction and graph matching. It was reported in [5] that the elastic graph matching process took 30 seconds on a SPARC station 10-512. Compared with their approach, our method is much faster and efficient.

Table 3: List of different prob sets.

Prob set	Gallery	Prob set size	Gallery size	Variations
Fb	Fa	1195	1196	Expression
Fc	Fa	194	1196	Illumination and camera
Dup I	Fa	722	1196	Time gap < 1 week
Dup II	Fa	234	1196	Time gap > 1 year

7. CONCLUSIONS

Mutual information theory has been successfully applied to select informative Gabor features for face recognition. To reduce the computation cost, the intrapersonal and extrapersonal difference spaces are defined. The Gabor features thus selected are nonredundant while carrying important information about the identity of face images. They are further enhanced in the nonlinear kernel space. Our algorithm has been tested extensively. The results on the whole FERET database also show that our algorithm achieves better performance on 3 test data sets than the top method in the competition—the elastic graph matching algorithm. Particularly, our method gives significantly better performance on the most difficult test set Dup II. Furthermore, our algorithm has advantage in computation efficiency since no graph matching process is needed.

Whilst we model features as binary random variables, the method could certainly be extended for continuous variables. However, as shown in Table 1, most of the feature distributions are non-Gaussian. As a result, a Gaussian mixture model may be needed to represent the distribution of features. When the random variables with multiple values are used, the selection process will require much more computation. The number of features to be selected is currently decided by experiments. A more advanced method is to use the information gain. If the gain by including a new feature is less than a threshold, we can say that the inclusion of new feature does not bring any more useful information. We are currently working on how to determine the threshold.

(a) (b) (c) (d) (e)

Figure 8: Examples of different probe images.

Table 4: FERET evaluation results for various face recognition algorithms.

Methods	Fb	Fc	Dup I	Dup II
PCA	83.4%	18.2%	40.8%	17.0%
PCA + Bayesian	94.8%	32.0%	57.6%	35.0%
LDA	96.1%	58.8%	47.2%	20.9%
Elastic graph matching	95.0%	82.0%	59.1%	52.1%
Variable Gabor features [27]	96.3%	69.6%	58.3%	47.4%
InfoGabor-GDA	96.9%	85.57%	55.54%	65.38%

REFERENCES

[1] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society of America A - Optics, Image Science, and Vision*, vol. 2, no. 7, pp. 1160–1169, 1985.

[2] K. Okajima, "Two-dimensional Gabor-type receptive field as derived by mutual information maximization," *Neural Networks*, vol. 11, no. 3, pp. 441–447, 1998.

[3] V. Kyrki, J.-K. Kamarainen, and H. Kälviäinen, "Simple Gabor feature space for invariant object recognition," *Pattern Recognition Letters*, vol. 25, no. 3, pp. 311–318, 2004.

[4] 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*, vol. 22, no. 10, pp. 1090–1104, 2000.

[5] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[6] K. Messer, J. Kittler, M. Sadeghi, et al., "Face authentication test on the BANCA database," in *Proceedings of 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 4, pp. 523–532, Cambridge, UK, August 2004.

[7] M. Lades, J. C. Vorbruggen, J. Buhmann, et al., "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 300–311, 1993.

[8] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[9] L. Shen and L. Bai, "Gabor feature based face recognition using Kernel methods," in *Proceedings of 6th IEEE International Conference on Automatic Face and Gesture Recognition(FGR '04)*, pp. 170–176, Seoul, South Korea, May 2004.

[10] I. R. Fasel, M. S. Bartlett, and J. R. Movellan, "A comparison of Gabor filter methods for automatic detection of facial landmarks," in *Proceedings of 5th IEEE International Conference on Automatic Face and Gesture Recognition(FGR '02)* , pp. 231–235, Washington, DC, USA, May 2002.

[11] D.-H. Liu, K.-M. Lam, and L.-S. Shen, "Optimal sampling of Gabor features for face recognition," *Pattern Recognition Letters*, vol. 25, no. 2, pp. 267–276, 2004.

[12] N. W. Campbell and B. T. Thomas, "Automatic selection of Gabor filters for pixel classification," in *Proceeding of 6th IEE International Conference on Image Processing and Its Applications(IPA '97)*, vol. 2, pp. 761–765, Dublin, Ireland, July 1997.

[13] Z. Sun, G. Bebis, and R. Miller, "Evaluationary Gabor filter optimization with application to vehicle detection," in *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM '03)*, pp. 307–314, Melbourne, Fla, USA, November 2003.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. 511–518, Kauai, Hawaii, USA, December 2001.

[15] L. Shen and L. Bai, "AdaBoost Gabor feature selection for classification," in *Proceeding of Image and Vision Computing Conference (IVCNZ '04)*, pp. 77–83, Akaroa, New Zealand, 2004.

[16] S. Z. Li and Z. Zhang, "FloatBoost learning and statistical face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1112–1123, 2004.

[17] G. D. Tourassi, E. D. Frederick, M. K. Markey, and C. E. Floyd Jr., "Application of the mutual information criterion for

feature selection in computer-aided diagnosis," *Medical Physics*, vol. 28, no. 12, pp. 2394–2402, 2001.
[18] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.
[19] T. P. Weldon, W. E. Higgins, and D. F. Dunn, "Efficient Gabor filter design for texture segmentation," *Pattern Recognition*, vol. 29, no. 12, pp. 2005–2015, 1996.
[20] V. Kruger and G. Sommer, "Gabor wavelet networks for efficient head pose estimation," *Image and Vision Computing*, vol. 20, no. 9-10, pp. 665–672, 2002.
[21] P. J. Phillips, "Support vector machines applied to face recognition," in *Proceedings of 1998 Conference on Advances in Neural Information Processing Systems II*, pp. 803–809, November 1999.
[22] B. Scholkopf, S. Mika, C. J. C. Burges, et al., "Input space versus feature space in Kernel-based methods," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, 1999.
[23] M.-H. Yang, "Kernel eigenfaces vs. Kernel fisherfaces: face recognition using Kernel methods," in *Proceedings of 5th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '02)*, pp. 215–220, Washington, DC, USA, May 2002.
[24] G. Baudat and F. Anouar, "Generalized discriminant analysis using a Kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.
[25] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
[26] M. Kendall, A. Stuart, and J. K. Ord, *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*, Edward Arnold, Paris, France, 1994.
[27] B. Kepenekci, F. B. Tek, and G. B. Akar, "Occluded face recognition based on Gabor wavelets," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '02)*, vol. 1, pp. 293–296, Rochester, NY, USA, September 2002.
[28] R. Beveridge and B. Draper, "Evaluation of Face Recognition Algorithms," 2003.

Linlin Shen received the B.Eng. and M.Eng. degrees in electronics engineering from Shanghai JiaoTong University, China, in 1997 and 2000, respectively. Currently, he is a Ph.D. student studying at the School of Computer Science and Information Technology, The University of Nottingham, UK. His research interests include face recognition, fingerprint recognition, kernel methods, boosting algorithm, computer vision, and medical image processing.

Li Bai is an Associated Professor in the School of Computer Science and Information Technology, The University of Nottingham, UK. She has a B.S. and an M.S. degree in mathematics, a Ph.D. degree in computer science. Her research interests are in the areas of pattern recognition, computer vision, and artificial intelligence techniques. She has been an academic referee for a number of journals and has published widely in international journals and conferences.