An Illumination Invariant Bimodal Method Employing Discriminant Features for Face Recognition

JiYing WU†, Student Member, QiuQi RUAN†, and Gaoyun AN†, Nonmembers

SUMMARY A novel bimodal method for face recognition under low-level lighting conditions is proposed. It fuses an enhanced gray level image and an illumination-invariant geometric image at the feature-level. To further improve the recognition performance under large variations in attributes such as poses and expressions, discriminant features are extracted from source images using the wavelet transform-based method. Features are adaptively fused to reconstruct the final face sample. Then FLD is used to generate a supervised discriminant space for the classification task. Experiments show that the bimodal method outperforms conventional methods under complex conditions.

key words: enhanced gray level image, illumination invariant image, discriminant feature extraction, image fusion

1. Introduction

Appearance-based models achieve good performance in face recognition. Among them PCA [1] and FLD [2] are two standard ones proposed for dimension reduction. FLD outperforms PCA significantly, because it uses Fisher’s Linear Discriminant rule to reduce the dimension while maintaining the discriminability of low-dimensional feature vector space. But FLD still achieves poor performance under low-level lighting conditions. Recognition rates of all these unimodal methods are negatively affected by the variations of pose, lighting and expression [3]. A robust face recognition system should fuse different kinds of information for enhanced recognition performance. Therefore image fusion is introduced for face recognition and multimodal methods are proposed.

There are three kinds of image fusion methods: pixel-level, feature-level and decision-level fusion [4]. In pixel-level fusion, a pixel’s gray value in the final face sample is determined by the average combination of pixels in source images. Then the fused sample is used for subspace analysis. In feature-level fusion, feature vectors computed from source images are fused to generate a feature vector. Face sample is reconstructed according to the vector. In decision-level fusion, recognition results of source images are fused to make a final decision using voting protocol.

In this paper, a bimodal method is proposed for face recognition under low-level lighting conditions and with large variations of pose and expression. The bimodal method fuses source images at the feature-level, which means the discriminant features are extracted and fused to generate the face sample. One of the source images is an enhanced gray level image which contains the holistic information, and the other is an illumination invariant image which contains the intrinsic geometric features. All illumination effects are removed in the second source, and the features extracted from it improve the face recognition performance under low-level lighting conditions. To further improve recognition performance with variations of pose and expression, the discriminant edge features are extracted by the wavelet transform-based method, which is geometric-feature-guided and contrast-enhanced. Then the face sample is reconstructed according to the fused discriminant features. Since sample labels are important pre-information for classification tasks in most cases, we further construct a supervised discriminant features space employing FLD in the fused face image space.

2. Face Recognition by Bimodal Method

2.1 Two Source Images

Both holistic image information and illumination invariant facial features are adopted in the bimodal method. One source image is a gray level image, which contains the holistic image information. Noisy pixels in it are smoothed by a partial differential equation (PDE), and the contrast of the whole image is enhanced by histogram equalization (HE). The other source image is illumination invariant. The infrared (NIR) image contains illumination invariant features; therefore, using it could improve recognition performance under low-level lighting conditions [5]. But NIR image are collected under different views, poses and imaging conditions with ordinary gray image. To be fused together, they should be registered firstly which is computationally expensive. Therefore NIR image is not adopted.

An illumination invariant image generated by a PDE is adopted in our bimodal method. This image is generated by geometrical computation of the same image with the enhanced gray image. The geometrical computation is implemented by PDE based on local geometry structure. So no registration is needed for the fusion of two images.

2.1.1 Enhanced Gray Level Image

Given the training set $X = \{x_1, x_2, \ldots, x_N\}$, where $x_i \in \mathbb{R}^{n \times 1}$ is the gray level distribution of the $i$th face image, and
\[ \hat{u}_i = \arg \min_{u_i} \int_{\Omega} |\nabla u_i| \, dx + \lambda_{L^2} \int |x_i - u_i| L^2 \, dx \]
\[ = \arg \min_{u_i} \int_{\Omega} |\nabla u_i| \, dx + \lambda_{L^2} \int (x_i - u_i)^2 \, dx \]  
(1)

where \( \lambda_{L^2} \) is the scale coefficient to control the relationship between fidelity term and smoothing term. TV-L^2 model is anisotropic diffused and it smoothes noisy pixels while preserving edge information. The denoised result of it is better than the conventional Gaussian filter. Then all information in \( \hat{u}_i \) is enhanced by HE. A flat histogram \( H \) with gray level \( K \) is created as:

\[ H = [1 \ 1 \ \cdots \ 1]_1 \times K \ast (n^2/K) \]  
(2)

With the histogram \( H \), a gray scale transformation \( T(\bullet) \) is chosen to minimize:

\[ \lfloor h_1(T(k)) - h_0(k) \rfloor \]  
(3)

where \( h_0(\bullet) \) is the cumulative histogram of \( \hat{u}_i \), \( h_1(\bullet) \) is the cumulative sum of \( H \) for all intensities \( k \). Generating the gray scale transformation \( T(\bullet) \), information in \( \hat{u}_i \) is enhanced:

\[ \hat{u}^*_i = T(\hat{u}_i) \]  
(4)

2.1.2 Illumination Invariant Image

The illumination effects in face image obey the Lambertian surface:

\[ x_i = \rho n^T s \]  
(5)

where \( \rho \) is the albedo (intrinsic features) of image, \( n \) is the normal direction of pixel, \( s \) is the lighting source. According to Eq. (5), the illumination invariant intrinsic features \( \rho \) are proportional to the illumination effects \( s \). So \( \rho \) could be generated by dividing the original image by the received light:

\[ \rho = x_i/n^T s \]  
(6)

where \( n^T s \) denotes the received light on image surface. In face image the intrinsic features, such as eyes and eyebrows, are represented as small-scale information. The illumination effects and shadows caused by large objects are large-scale information. TV-L^1 (Total variation with L^1-norm constrained) is an image decomposition mode [7]. It could generate a large-scale representation \( u_i \) and preserve all the small scale information in the left part:

\[ u_i = \arg \min_{u_i} \int_{\Omega} |\nabla u_i| \, dx + \lambda_{L^1} \int |x_i - u_i| L^1 \, dx \]  
(7)

where \( \lambda_{L^1} \) is determined by the size of sample, \( u_i \) contains the lighting effects, and the intrinsic features are preserved in \( x_i - u_i \). Therefore, an illumination invariant image could be computed:

\[ v_i = x_i/u_i \]  
(8)

2.2 Discriminant Features Extraction

The information in two source images is then fused at the feature-level. Discriminant features for recognition are extracted from source images by wavelet transform-based method using Mallat’s wavelet [8]. Two source images are denoted as \( J(x,y) \) and \( V(x,y) \) for gray image and illumination invariant image, respectively, in the fusion procedure. The wavelet feature extraction procedure is:

1. Constructing the smoothing filter \( \theta(x,y) \). \( \theta(x,y) \) is chosen as the Gaussian filter:

\[ \theta(x,y) = \frac{1}{2\pi s^2} \exp\left(-\frac{x^2 + y^2}{2s^2}\right) \]  
(9)

where \( s \) is convolving scale of the Gaussian filter.

2. Computing the first order derivative of the convolved result of \( \theta(x,y) \):

\[ \psi^1(x,y) = \frac{\partial}{\partial x} \theta(x,y), \quad \psi^2(x,y) = \frac{\partial}{\partial y} \theta(x,y) \]  
(10)

\[ \psi^3(x,y) \) and \( \psi^4(x,y) \) are used to detect the gradient vector pixels in the image.

3. Generating the wavelets \( W^1_J(x,y) \) and \( W^2_J(x,y) \) with scale \( s \). The wavelets for \( J(x,y) \) and \( V(x,y) \) are:

\[ W^1_J(x,y) = J \ast \psi^1_J(x,y), \quad W^2_J(x,y) = J \ast \psi^2_J(x,y) \]  
(11)

\[ W^1_V(x,y) = V \ast \psi^1_J(x,y), \quad W^2_V(x,y) = V \ast \psi^2_J(x,y) \]  
(11)

4. Computing the modulus \( M_J(x,y) \) which is used to extract the discriminate image features. The wavelet transform of the image \( J(x,y) \) is:

\[ \begin{pmatrix} W^1_J(x,y) \\ W^2_J(x,y) \end{pmatrix} = s \begin{pmatrix} \frac{\partial}{\partial x} (J \ast \theta_J(x,y)) \\ \frac{\partial}{\partial y} (J \ast \theta_J(x,y)) \end{pmatrix} = s \nabla (J \ast \theta_J(x,y)) \]  
(12)

The transform for image \( V(x,y) \) has the same form with Eq. (10). So the wavelet modulus is:

\[ M_J(x,y) = \sqrt{|W^1_J(x,y)|^2 + |W^2_J(x,y)|^2} \]  
(13)

Then the modulus of \( J(x,y) \) and \( V(x,y) \) are computed as \( M_{J_J}(x,y) \) and \( M_{V_J}(x,y) \), which are the wavelet magnitudes of two source images.

2.3 Adaptive Feature Fusion and Image Reconstruction

The final modulus \( M_J(x,y) \) are adaptively determined by \( M_{J_J}(x,y) \) and \( M_{V_J}(x,y) \) to generate the fused feature. To maintain the discriminant edge features, \( M_{V_J}(x,y) \) should be all contained in \( M_J(x,y) \), since \( V(x,y) \) is the normalized result of image. All large-scaled information is removed and \( V(x,y) \) contains only the edge and texture information:

\[ M_J(x,y) = \begin{cases} c_J \cdot \max(M_{J_J}(x,y), \ M_{V_J}(x,y)), & \text{if } M_{J_J}(x,y) \text{ and } M_{V_J}(x,y) \text{ are maximas} \\ c_J \cdot M_{V_J}(x,y), & \text{else} \end{cases} \]  
(14)
where \( c_s \) is used to stretch the gradient maxima and \( c_s \geq 1 \) in different scale \( s \). When \( M_{I_s}(x,y) \) is maxima \( I(x,y) \) is an edge pixel, and then the edge information in both source images is combined; else only the information in \( V(x,y) \) is contained in the final modulus.

Lastly the face sample is generated by iterative method proposed by Mallat [9]. The reconstructed face sample is:

\[
Y(x, y) = W^{-1}(M_s(x, y))
\]

2.4 FLD

Subspace approach is used to extract features in the final face sample for classifier. Subspace approach projects the high-dimensional vector space into a low-dimensional feature vector space. FLD is optimal from a discrimination standpoint. The image space reconstructed from the fused discriminant features is \( Y(x, y) \). FLD selects the optimal projection \( W \) by:

\[
W_{\text{opt}} = \arg\max_W \frac{|W^T S_B W|}{|W^T S_W W|} = [w_1 w_2 \cdots w_m]
\]

where \( \{w_i | i = 1, 2, \ldots m\} \) is the set of generalized eigenvectors of \( S_B \) and \( S_W \) corresponding to the \( m \) largest eigenvectors \( \{\lambda_i | i = 1, 2, \ldots m\} \). In \( Y(x, y) \), the between-class scatter matrix is:

\[
S_B = \sum_{i=1}^{c} N_i (\mu_i - \mu)(\mu_i - \mu)^T
\]

and the with-class scatter matrix is:

\[
S_W = \sum_{i=1}^{c} \sum_{y_k \in \mathcal{V}_i} (y_k - \mu_i)(y_k - \mu_i)^T
\]

where \( \mu_i \) is the mean image of \( Y(x, y) \) and \( N_i \) is number of samples in \( Y(x, y) \). The classify feature is computed as:

\[
z = W_{\text{opt}}^T V(Y(x, y))
\]

where \( V(\cdot) \) stands for formatting a matrix to a row vector by concatenating its rows or columns of sample \( Y(x, y) \). The vector \( z \) is used for the classifier.

3. Experimental Results

The CMU PIE face database is used in the following experiments, which contains 41368 samples of 68 subjects with different pose, illumination and expression. In the experiments, all the samples are cropped and rotated to \( 32 \times 32 \) images according to the coordinates of two eyes and have 256 gray levels. The dimension of feature space is 67 for FLD. A random subset with \( n \) samples per subject was chosen to form the training set, and the rest samples of CMU PIE database form the testing set. For each chosen \( n \), we do separate experiment 30 times and average all the results.

Table 1 illustrates the accuracy recognition rates of different approaches. We choose HE and TVQI (Total variation-based quotient image) methods to be compared with our bimodal method. HE is the most widely used methods and TVQI is proved to achieve the best performance under low-level lighting conditions recently [9].

In order to classify feature \( z \) of sample \( Y(x, y) \), the 1-nearest neighbor classifier is adopted with four famous distances (L1: Absolute Value Distance, L2: Euclidean Distance, Cos: Cosine Distance and Md: Mahalanobis Distance) as similarity measurement for classifier. As there is only one sample per subject/class in the training set, the same covariance matrix is used for all classes for Mahalanobis Distance and \( \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_m^2) \).

3.1 Experiment in Small Scale Dataset

Firstly, to compare our work with TVQI [9], the same experiment with [9] is repeated. The frontal face images of 68 subjects in the CMU database, each with 21 different illuminations, are used for evaluation. There are total 1428 samples in chosen dataset, and it is a small scale dataset.

The rates shown in Table 1 are consistent with the results in [9]. Both TVQI+FLD [9] and our approach could achieve the similar performance in a small scale dataset.

3.2 Experiment in Large Scale Dataset

To further evaluate performances of methods with a large scale dataset, we choose all the 170 near frontal samples of each subject under different illumination, lighting and expression to form our dataset, which is shown in Fig. 1. Therefore there are total 11554 samples (6 original samples are corrupted and lost) of 68 subjects in the chosen dataset, it is a large-scale dataset.

Table 2 illustrates the accuracy recognition rates at rank 1 with 10 samples per subject formatting the training set. If the max accuracy recognition rates of various
Table 2  The accuracy recognition rates of different approaches.

|       | FLD [2] | IIE+FLD | TVQI+FLD [9] | Our approach |
|-------|---------|---------|--------------|--------------|
| L1    | 69.12   | 78.43   | 70.57        | 82.88        |
| L2    | 70.77   | 80.23   | 72.40        | 83.45        |
| Cos   | 73.10   | 80.78   | 73.43        | 86.29        |
| Md    | 70.82   | 80.27   | 72.45        | 84.70        |

**Fig. 2**  The error rate with different number of samples per subject in the training set on the CMU PIE face database.

methods are considered, the bimodal method outperforms FLD [2], FLD with HE and TVQI with FLD [10] for all testing sets.

FLD achieves limited recognition performance under low-level lighting conditions. HE enhances the image information destroyed by the lighting effects, and it has a better recognition rate compared with FLD. TVQI works poorly in this experiment because it contains only small-scaled facial features. The small-scaled features represent the small intrinsic details in face image, such as eyes, nose etc.; in the contrary, the large-scaled features represent the large intrinsic features, illumination effects [10]. It is already proved the small-scaled features could not discriminate the face samples in large scale face databases [10]. The bimodal method achieves the best performance since the most discriminant features from two sources are fused and sufficient features are extracted to assist recognition.

Figure 2 illustrates the error recognition rate with different number of samples per subject in the training set. According to the rates shown in Table 2, all the methods achieve best performance with Cos distance as similarity measurement. Therefore, all the approaches in Fig. 2 are compared with Cos distance. The experimental results confirm that our bimodal method could outperform FLD, HE with FLD, and TVQI with FLD for all testing sets in the large scale database. It achieves the lowest error recognition rates under most testing conditions, except for 2 samples per subject.

### 4. Conclusions

A bimodal method for face recognition under low-level lighting conditions is proposed. The illumination invariant image is fused with a gray level image at the feature level. The discriminant features are extracted from source images to further improve recognition performance under other variations, such as pose and expression. Then FLD is used to generate a supervised discriminant space for classification. According to the experiments on CMU PIE datasets, our bimodal method outperforms conventional methods. The feature-level fusion method preserves the most discriminant features and largely improves the recognition performance under complex conditions.

**Acknowledgments**

This paper was partially supported by the National Natural Science Foundation of China (No. 60672062), the 973 Program of China (No. 2004CB318005), and the Technological Innovation Fund of Excellent Doctorial Candidate of Beijing Jiaotong University (No.48026).

**References**

[1] M. Kirby and L. Sirovich, “Application of the Karhunen-Loeve procedure for the characterization of human faces,” IEEE Trans. Pattern Anal. Mach. Intell., vol.12, no.1, pp.103–108, 1990.

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.7, pp.711–720, 1997.

[3] J.G. Wang, W.Y. Yau, A. Suwandy, and E. Sung, “Person recognition by fusing palmprint and palm vein images based on “Laplacianpalm” representation,” Pattern Recognit., vol.41, pp.1514–1527, 2008.

[4] A.A. Ross, K. Nandakumar, and A.K. Jain, Handbook of multimodality, Springer, 2006.

[5] S.Z. Li, R.F. Chu, S.C. Liao, and L. Zhang, “Illumination invariant face recognition using near-infrared images,” IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.4, pp.627–639, 2007.

[6] S. Mallat and S. Zhong, “Characterization of signals from multiscale edges,” IEEE Trans. Pattern Anal. Mach. Intell., vol.14, no.7, pp.710–732, 1992.

[7] X. Xie, W. Zhang, J. Lai, and P. Yuen, “Face illumination normalization on large and small scale features,” Proc. CVPR08, pp.1–8, Alaska, USA, 2008.