HUMAN FACE CLASSIFICATION BASED ON LOCALIZED BLUR DESCRIPTORS

Abdul Adeel Mohammed, Q.M. Jonathan Wu, Maher A. Sid-Ahmed

Department of Electrical Engineering
University of Windsor, 401 Sunset Ave, Windsor, ON, Canada

ABSTRACT

In our proposed work localized patch based geometric blur point descriptors are accumulated to generate a global similarity matrix for every pair of *focal* and *query* image. A *focal* image is a randomly selected template image that represents each class and a *query* image symbolizes all other images that belong to the same class. The similarity matrix is dimensionally reduced using the proposed bidirectional 2-dimensional principal component analysis technique to generate distinctive feature sets. These feature sets are used for training and testing an extreme learning machine classifier. The proposed face recognition structure handles variations in head positions, lighting conditions, facial expressions and cluttered background by exclusively matching template and query images. Extensive experiments are performed using challenging face databases and significant improvements in recognition accuracy were achieved.

***Index Terms*—** Face recognition, geometric blur descriptors, principal component analysis, extreme learning machine.

1. INTRODUCTION

Face recognition is a standard machine learning problem wherein given a set of labeled and unlabeled face image, i.e., the gallery set and probe set, we seek to ascertain the identity of every person from amongst the collection of probe images. Geometric feature based face recognition is achieved by evaluating attributes such as distances and angles between various facial features such as eyes, nose, mouth and chin [1]. Feature-based methods depend upon feature extraction and measurement process and research has shown that face recognition methods based on finding local image features and inferring identity using geometric relations is often ineffective [2]. Appearance-based methods have emerged as fitting alternatives that use low-dimensional representation of face images to perform recognition. A human face is treated as a two-dimensional intensity variation pattern and recognition is established through detection and matching of statistical properties [3]. In this work we emphasize that local feature based face recognition can indeed achieve superior recognition in comparison to conventional appearance-based recognition schemes without explicitly generating a list of distance and angles attributes amongst diverse facial elements.

To improve precision and to deal with high image dimensionality, Principal Component Analysis (PCA) and other dimensionality reduction techniques have been proposed. Kirby *et al.* [4] represented human faces as a linear combination of weighted eigenvectors using PCA. In [5], authors used Linear Discriminant Analysis (LDA) to maximize the ratio of between-class scatter matrix and the within-class scatter matrix for improved face recognition. The scatter matrices in LDA are evaluated based on the assumption that the samples in each class satisfy the Gaussian distribution. Fukunaga [6] proposed a Non-parametric Discriminant Analysis (NDA) in cases of non-normal distribution constraints by defining a new between-class scatter matrix. Loog and Duin [7] proposed a heteroscedastic LDA (HLDA) to deal with data that contains classes with unequal covariance matrices. Researchers have also proposed the use of Support Vector Machine (SVM) [8] and recently Ksantini *et al.* [9] proposed a novel Bayesian Logistic Discriminant (BLD) representation that deals with both the normality and heteroscedasticity problems.

In our proposed scheme a human face is represented by a set of points sampled from the contours of a facial image. These pixel locations are sampled from the output of an edge detector and it is worth mentioning that there is nothing special about these points and that they do not represent key points found using a Harris/Forstner corner detectors or Lowe's scale-space extrema of a Laplacian of Gaussian operator. Earlier works have shown the benefit of using filter-based patch features for shape or texture based recognition since corresponding points on two shapes generate similar local descriptors. In our proposed recognition scheme we use geometric blur descriptors [10] to generate localized patch based similarities, these similarity functions are accrued to generate a similarity matrix between the *focal* and the *query* image. Dimensionality of the similarity matrix is reduced using our proposed Bidirectional two-dimensional Principal Component Analysis (B2DPCA) scheme [11]. Finally reduced dimension vectors are used for training and testing an Extreme Learning Machine (ELM) classifier.

2. GEOMETRIC BLUR DESCRIPTOR

Recognition based on key points alone sacrifices the shape information available in smooth segments of object contours, thus approaches based on extracting edge points are more universal. The signal and the blur function are carefully designed in order to produce distinguishing features. A sparse signal indicates the presence of some interesting feature points such as an image edge whereas a blur function determines the nature of expected geometric distortion. An edge detector is used to generate oriented edge responses and in our work Martin's [12] oriented boundary edge detector was employed. Four oriented sparse energy channels are simultaneously used to compute geometric blur descriptors.

Geometric blur is used to estimate a robust similarity measure that requires computation of an average signal over its smoothed versions, blurred by convolving the signal with a spatially varying kernel function. In other words, geometric blur descriptor blurs the signal over the range of acceptable transformations (small affine transformations). Therefore, signal points that are far from the center are blurred more since they have the opportunity to move more. We exploit this property and subsample the geometric blur as shown in Figure 1. The objective is to generate discriminative information by using an extended patch of signal and thus provide greater robustness to distortions. Once such a blur is evaluated correlation can accurately identify similar pairs of local patches.

978-1-4577-1303-3/11/$26.00 ©2011 IEEE

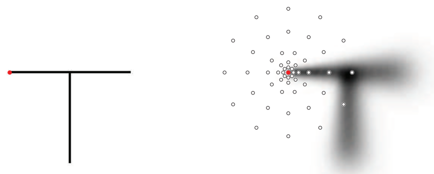


Fig. 1. A sparse signal *S* (left) and the geometric blur of *S* around the feature points marked in red [10].

A spatially varying Gaussian kernel function is used to evaluate the geometric blur. If every oriented edge energy channel is labeled as the signal φ, we compute its blurred version, φ_σ by convolving it with a Gaussian kernel of standard deviation σ. The geometric blur around a feature point located at ϕ_0 is calculated using Equ. 1.

$$\varphi_\sigma = \varphi * \Omega_\sigma, \quad B_{\phi_0}(\phi) = \varphi_{\mu|\phi|+\kappa}(\phi_0 - \phi) \tag{1}$$

μ and κ are constants that determine the amount of geometric blur. It is perceived that under affine transformation that fixes a feature point, the distance a segment of the signal moves is directly proportional to its distance from the feature point. The signal $B_{\phi_0}(\phi)$ is sampled for a sparse set of points $\phi = \varphi_i$ and φ_σ is computed for distinct values of $\sigma = \mu|\varphi_i| + \kappa$. The feature descriptor at each point is the concatenation of the subsampled geometric blur descriptors at that point in the four channels. The degree of correspondence is measured between each *focal* and *query* image to generate a 'global' similarity matrix whose dimensionality is reduced.

3. BIDIRECTIONAL TWO DIMENSIONAL PRINCIPAL COMPONENT ANALYSIS

Kirby and Sirovich [4] used PCA for enhanced representation of face images, however PCA fails to capture even a minor variance unless it is explicitly accounted in the training data. Wiskott *et al.* [13] proposed a bunch graph matching technique to overcome limitations and weakness of linear PCA. In [14] Yang *et al.* proposed two dimensional PCA (2DPCA) for image representation. As opposed to PCA, 2DPCA is based on 2D matrices rather than 1D vectors. Therefore, image matrix is not vectorized and an image covariance matrix is directly computed using the original image matrices.

Let X denote a q dimensional unitary column vector. To project a pxq image matrix A to X; linear transformation $Y = AX$ is used which results in a p dimensional projected vector Y. The total scatter of the projected samples is determined to measure the discriminatory power of the projection vector X. The total scatter is characterized by the trace of S_x (covariance matrix of the projected feature vectors), $J(X) = tr(S_x)$, where $tr()$ represents the trace of S_x.

$$tr(S_x) = X^T[E(A - EA)^T(A - EA)]X \tag{2}$$

$G_t = E[(A - EA)^T(A - EA)]$ is a nonnegative qxq image covariance matrix. If there are M training samples, the α^{th} image sample is denoted by pxq matrix A_α.

$$G_t = \frac{1}{M}\sum_{\alpha=1}^{M}(A_\alpha - A)^T(A_\alpha - A) \tag{3}$$

$$J(X) = X^T G_t X \tag{4}$$

A represents an average image of all the training samples. The unitary vector X_{opt} that maximizes the generalized total scatter criterion $J(X)$ is called the optimal projection axes. X_{opt} represents a collection of d orthonormal eigen vectors $X_1, X_2, ..., X_d$ of G_t corresponding to d largest eigen values. Hence dimensionality of every image A_α is reduced by post multiplying and pre-multiplying it by the optimal projection axes as $X_{opt}^T A_\alpha X_{opt}$.

A limitation of 2DPCA based recognition is that it operates along row directions only. Our dimensionality reduction algorithm works along the row and column directions independently of one another in order to better preserve the neighborhood relationship and to generate distinctive feature sets. Our proposed technique closely follows the work of [14] and generates an image covariance matrix $G_{t\alpha}$ and optimizes it. Once optimal projection axes $X_{opt\alpha}$ is calculated, dimensionality of every image A_α is reduced along its columns to generate new image sets A_β using equation (5). The process is repeated to further reduce row dimension of the newly generated image sets by analyzing image covariance matrix $G_{t\beta}$ and optimal projection axes $X_{opt\beta}$ and finally pre-multiplying every new image A_β with $X_{opt\beta}^T$ using equation (6).

$$A_\beta = A_\alpha X_{opt\alpha} \tag{5}$$

$$A_\Theta = X_{opt\beta}^T A_\beta \tag{6}$$

4. EXTREME LEARNING MACHINE

Feedforward neural networks are ideal classifiers for non-linear mappings that utilize gradient descent approach for weights and bias optimization. Important factors that influence the performance of traditional feedforward neural learning algorithm like Back-Propagation (BP) include:

- A small value of learning parameter ρ causes the learning algorithm to converge **slowly** whereas a higher value leads to **instability and divergence to local minima**.
- Neural network may be **over-trained** using BP and obtain **inferior generalization performance**.
- Gradient descent based learning is an extremely **time consuming** process for most applications.

To overcome innate shortcomings of traditional learning techniques Huang *et al.* [15] proposed ELM to train a Single-hidden Layer Feedforward neural Network (SLFNN). A random selection of input weights and the hidden layer biases transforms the SLFNN into a linear system. Consequently, the output weights can be analytically determined through a simple generalized inverse operation of the hidden layer output matrices. In ELM an infinitely differentiable hidden layer activation function facilitates random assignment of input weights and hidden layer biases. Consider a collection of N distinct samples (x_i, t_i) where $x_i = [x_{i1}, x_{i2}, ...x_{in}]^T \in \Re^n$ and $t_i = [t_{i1}, t_{i2}, ...t_{im}]^T \in \Re^m$, an ELM with L hidden nodes and an activation function $\xi(x)$ is modeled as:

$$\sum_{i=1}^{L}\gamma_i\xi_i(x_n) = \sum_{i=1}^{L}\gamma_i\xi_i(w_i x_n + b_i) = t_n, n = 1, 2, ...N, \tag{7}$$

where $w_i = [w_{i1}, w_{i2}, ...w_{in}]^T$ and $\gamma_i = [\gamma_{i1}, \gamma_{i2}, ...\gamma_{iL}]^T$ represent input and hidden layer weight vectors respectively. ELM reliably approximates N samples with minimum error. Equation (7) is modified as $\delta\gamma = \tau$, $\delta = (w_1, ..., w_L, b_1, ..., b_L\ , x_1, ..., x_N)$, such that i^{th} column of δ is the output of i^{th} hidden node with respect to inputs $x_1, x_2, ...x_N$. If the activation function $\xi(x)$ is infinitely differentiable, it is proved that the number of hidden nodes satisfy $L \ll N$. Training of ELM requires minimization of an error function E.

$$E = \sum_{n=1}^{N}(\sum_{i=1}^{L}\gamma_i\xi_i(w_ix_n + b_i) - t_n)^2 \Rightarrow E = \|\delta\gamma - \tau\|. \quad (8)$$

In classical neural networks δ is determined using gradient descent optimization wherein the input weights w_i, hidden layer weights γ_i and bias parameters b_i are iteratively tuned with a learning parameter ρ. To avoid instability and divergence to local minima, ELM incorporates a minimum norm least-square solution. Hence, the problem is transformed to a new domain and an optimal solution in the simplified domain is evaluated using the least-square solution. Instead of tuning the entire network parameters, input weights and bias parameters are randomly allocated and the problem is curtailed to the least-square solution of $\delta\gamma = \tau$. The hidden layer output matrix δ is a non-square matrix and the norm least-square solution reduces to $\gamma = \delta^*\tau$ (δ^* is the moore-penrose generalized inverse of δ). An infinitely small training error is achieved using the above model since it represents a least-square solution of the linear system.

$$\|\delta\gamma - \tau\| = \|\delta\delta^*\tau - \tau\| \equiv min_\gamma\|\delta\gamma - \tau\|. \quad (9)$$

5. PROPOSED ALGORITHM

Our proposed architecture is based on local patch-based shape and texture features that are invariant and robust to changes in scale, translation and affine deformations. The basic outline of our discriminative approach for face recognition is as follows: (1) randomly select a *focal* image, training and testing set for every subject class, (2) for every image select a subset of interest points and localized patches surrounding it. (3) Compute fixed-length feature vectors for all interest points from each image and (4) define a similarity metric to generate a similarity matrix between sets of images. (5) Reduce the dimensionality of the similarity matrices and generate discriminative feature vectors, and (6) finally, use the feature vectors to train and test an ELM classifier and determine system performance. Schematic block diagram of our proposed face proposed recognition algorithm is shown in Figure 2.

All images are converted to 8-bit gray level image and one image from each class is randomly selected as a *focal* image and the remaining images are divided into training and testing set. 50-55% of images of each subject are used as prototypes and the remaining images are used to assess accuracy. We assume no prior information about location, view point, background and/or image acquisition constraints. Instead of using direct intensity values to extract features we calculate oriented edge energy channels. In our proposed face recognition algorithm we utilize Martin's boundary edge detector [12] to generate 4 sparse oriented channels for every *focal*, train and test image. A sample face image and its respective oriented edge energy channels is shown in Figure 3.

350 sampled edge points from every image are selected and features are calculated by centering the patches of fixed scale and orientation at the selected points. A patch radius of 50 pixels was used to sample 51 points for every marked feature point as shown in Fig. 1. The sampled geometric descriptors computed at that point in each of the 4 oriented edge energy channels are concatenated to generate a feature descriptor of 204 dimensions. Finally, the feature descriptors are normalized to ensure that the L_2 norm is unity. Generation of a feature descriptor representing various important feature points

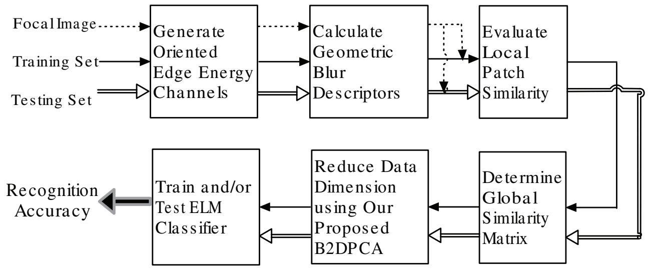


Fig. 2. Schematic diagram of our face recognition algorithm.

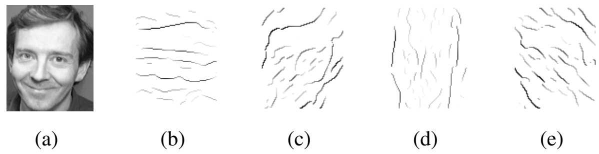


Fig. 3. (a) Original image (b-d) Oriented edge energy channels.

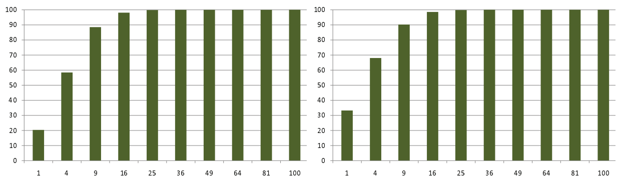
of an image allows us to incorporate varying geometric attributes of an object and to achieve better categorization.

The degree of correspondence between images is measured with respect to the similarity of feature points to their corresponding feature points. The similarity between geometric blur descriptors at respective local patches is evaluated using normalized cross correlation function. The local similarity patches between each *focal* and *query* image are assembled to generate a 'global' similarity matrix of dimension 350x350 between respective image sets. For real-time, accurate and efficient processing dimensionality of the generated similarity matrix is reduced. B2DPCA is applied to generate unique feature sets and minimize computational complexity of our framework. In our proposed work features are extracted by initially reducing dimension of initial feature matrix, i.e., similarity matrix along its columns, called as intermediate features. Later dimensionality of intermediate features is reduced along its rows so as to generate final feature sets. The B2DPCA approach used in this paper preserves information between adjoining pixels and generates distinctive feature vectors. An ELM classifier is trained and tested using labeled B2DPCA feature vectors to ascertain accuracy.

6. RESULTS AND EXPERIMENTS

All Images are transformed from RGB to gray level and local patch similarity is evaluated for every unique combination of *focal*, train and test image. Hence, a 'global' similarity matrix is calculated according to the method described in Section 5 and dimensionally reduced using B2DPCA. Dimensionally reduced features are vectorized, trained and tested using an ELM classifier with 50 hidden neurons. 100 experiments are conducted for each database and average results are calculated.

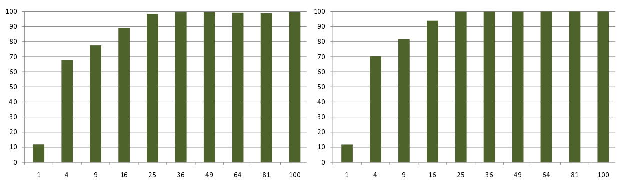
Two major factors that influence recognition rates are the number of principal components and the number of hidden layers neurons in the ELM structure. Accuracy analysis with respect to a change in feature size is critical for judging the competence of a recognition system. Figure 4 shows the average recognition rates at varying number of principal components for FERET and Yale database. Experimental results show that accuracy is directly proportional to the number of principal components. This represents a significant departure from traditional dimensionality reduction techniques where increasing the number of principal components do not necessarily guarantee improved detection accuracy. The proportional relationship is attributed to the use of localized patch information for face recognition. It should be noted that for simplicity the number of principal components are taken as integer squares due to application of B2DPCA along row and column directions respectively. Figure 4 vividly demonstrates that our proposed face recognition algorithm can deal with challenging situations: view point, facial expression, lighting and shadowing , gender, age etc.



(a) FERET (b) Yale

Fig. 4. Recognition rate (y-axis) vs. principal components (x-axis).

Experiments are also carried out by varying the number of hidden neurons, however, minor variations in accuracy were observed, as indicated by the recognition rates for Georgia Tech database in Figure 5. Based on experimental evidence it was established that a minimum of 50 hidden layer neurons ensure that the recognition rate always increases as the number of principal components are increased. Results corroborate that our proposed method is robust against changes in scale, orientation and lighting conditions. Therefore, based on desired accuracy and computational complexity we can adaptively select the number of principal components and the number of hidden layer neurons.



(a) 50 Neurons (b) 100 Neurons

Fig. 5. Recognition rate (y-axis) vs. principal components (x-axis).

To emphasize superiority of our proposed face recognition algorithm to other existing approaches we have compared the results obtained for ORL and AR face database against BLD, SVM, LDA, NDA and HLDA. Results in Table 1 highlight the improved performance of our proposed face recognition algorithm in comparison to existing modern techniques.

Table 1. Comparative results for ORL and AR database.

Database	Proposed	BLD	SVM	LDA	NDA	HLDA
ORL	**94.13**	91.9	90.5	87	85.2	79.1
AR	**89.34**	81.2	78.9	73.5	73.5	67

7. CONCLUSION

An efficient human face recognition technique based on localized patch based geometric blur descriptors is proposed. Geometric blur descriptors generate robust patch based similarity matrices between image sets. B2DPCA is employed to evolve distinctive feature sets which were trained and tested using a fast and accurate classifier. Experimental results substantiate our claim that the proposed method achieves improved recognition rate and ensures that a direct relationship exists between accuracy and the number of principal components. Furthermore, it enhances the user confidence and flexibility towards achieving desired accuracy based on recognition requirements.

8. REFERENCES

[1] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, and N. Otsu, Face Recognition System using Local Autocorrelations and Multiscale Integration, in Trans. on PAMI. IEEE, 1996, vol. 18, pp. 1024-1028.

[2] R. Brunelli and T. Poggio, Face Recognition: Features vs Templates, in Trans. on PAMI. IEEE, 1993, vol. 15, pp. 1042-1053.

[3] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, Face recognition: A Literature Survey, ACM Computing Surveys, 2003, vol. 35, pp. 399-458.

[4] M. Kirby and L. Sirovich, Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces, in Trans. on PAMI. IEEE, 1990, vol. 12, pp. 103-108.

[5] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection, in Trans. on PAMI. IEEE, 1997, vol. 19, pp. 711-720.

[6] K. Fukunaga, Introduction to Statistical Pattern Recognition, Second ed., Academic Press, NewYork, 2000.

[7] M. Loog, and R.P.W. Duin, Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion, in Trans. on PAMI. IEEE, 2004, vol. 26, pp. 732-739.

[8] B. Heisele, P. Ho, and T. Poggio, Face Recognition with Support Vector Machines: Global Versus Component-based Approach, in Proc. of ICCV, 2001, pp. 688-694.

[9] R. Ksantini, B. Boufama, D. Ziou and B. Colin, A Novel Bayesian Logistic Discriminant Model: An Application to Face Recognition, Pattern Recognition, 2010, vol. 43, pp. 1421-1430.

[10] A. Berg, T. Berg, and J. Malik, Shape Matching and Object Recognition using Low Distoriton Correspondence, in Proc. of CVPR, 2005, pp. 26-33.

[11] A.A. Mohammed, R.Minhas, Q.M. Jonathan Wu and M.A. Sid-Ahmed, Human Face Recognition based on Multidimensional PCA and Extreme Learning Machine, in Pattern Recognition, 2011, doi:10.1016/j.patcog.2011.03.013

[12] D. Martin, C. Fowlkes. and J. Malik, Learning to Detect Natural Iamge Boundaries using Local Brightness, Color and Texture Cues, in Trans. on PAMI. IEEE, 2004, vol. 26, pp. 530-549.

[13] L. Wiskott, J.M. Fellus, N. Kruger, and C. VonDerMalsburg, Face Recognition by Elastic Bunch Graph Matching, in Trans. on PAMI. IEEE, 1997, vol. 19, pp. 775-779.

[14] J. Yang, D. Zhang, F. Frangi and J.Y. Yang, Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition, in Trans. on PAMI. IEEE, 2004, vol. 26, pp. 131-137.

[15] G. B. Huang, Q.Y. Zhu, and C.K. Siew, Extreme Learning Machine: Theory and Applications, Neurocomputing, 2005, vol. 70, pp. 489-501.