FLEXIBLE X-Y PATCHES FOR FACE RECOGNITION

Ming Liu[1*], *Shuicheng Yan*[2], *Yun Fu*[1], *and Thomas S. Huang*[1]

[1]Beckman Institute, Uni. of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2]Dept. of ECE, National University of Singapore, Singapore
{mingliu1, yunfu2, huang}@ifp.uiuc.edu, eleyans@nus.edu.sg

ABSTRACT

In this paper, illuminated by the great success of Universal Background Modeling (UBM) for speech/speaker recognition, we present a new algorithm for face recognition. On the one hand, we encode each face image as an ensemble of X-Y patches, which integrate both local appearance and shape information. These X-Y patch representation provides the possibility to compare two spatially different patches, and consequently alleviates the requirement of exact pixel-wise alignment. On the other hand, we train the UBM based on the X-Y patches from the images of differen subjects, and then automatically adapt the UBM for specific subject, and finally face recognition is conducted by comparing the ratio of the likelihoods from the model for specific subject and UBM. UBM elicits the algorithmic robustness to image occlusion since the occluded patches may not contribute evidence to any subjects. Comparison experiments with the state-of-the-art subspace learning algorithms, on the popular CMU PIE face database and with varieties of configurations, demonstrate that our proposed algorithm brings significant improvement in face recognition accuracy, and also show the algorithmic robustness to image occlusions.

Index Terms— Appearance, Shape, X-Y Patches, Face Recognition.

1. INTRODUCTION

Face recognition has been an active research topic for several decades. Owing to the wide applications in biometrics, Human Computer Interface (HCI), and many other vision related fields, many algorithms have been proposed for providing different solutions. These algorithms can be roughly divided into two categories: appearance based and model based algorithms.

The appearance based category can be further divided into holistic and part-based algorithms. For holistic algorithms, the facial image is treated as a concatenated vector, and often dimensionality reduction techniques [6, 17], such as Principal Component Analysis (PCA) [15], Linear Discriminant Analysis (LDA) [4] and Independent Component Analysis (ICA) [10], are utilized for feature extraction before formal classification. PCA/LDA based algorithms consider the whole face image as a single feature vector, and hence implicitly assume the pixel-wise alignment between the probe and gallery images. However, pixel-wise alignment is often difficult or even impossible to achieve in real scenarios, especially for fully automatic systems. It brings the robustness issue of the traditional PCA/LDA based algorithms. Part-based algorithms instead work on image parts corresponding to specific facial features such as eyes, nose tip, mouse corners. Separate appearance models are learnt for different parts, and the final classification is conducted by fusing the outputs from all the parts [18]. In addition to the appearance of these parts, the shape information for these parts may also be taken into account for the classification. Wiskott et al. [16] proposed to represent human face as a undirected graph with nodes located at the fiducial points (eyes, nose tip, mouth corners, etc.), and the edges connecting the nodes are used to encode the geometric constraints of the human face. Heisele et. al [8] presented an algorithm to locate facial components, and them concatenate them into a long feature vector followed by Support Vector Machine (SVM) for classification.

In this paper, we present a new flexible representation, called X-Y Patches, for encoding the human faces. X-Y patches are localized and include both local appearance and relative coordinates information. In this sense, X-Y patch is a joint appearance and shape descriptor. The human face is then represented as a bag of X-Y patches. As the x-y coordinates are included in the patches, the comparison of the descriptors from different positions can be conducted in a flexible manner: on the one hand, two patches with the same appearance yet slight mismatch in spatial domain may be considered relatively similar; and on the other hand, two patches with the same appearance yet far way in spatial domain will be considered different. In this sense, the proposed descriptor alleviates the requirement of pixel-wise alignment and is robust to small spatial misalignment problem, which is often encountered in real applications. Instead of localizing the facial features, we adopt dense sampling scheme to extract the local X-Y patches from the facial images.

Based on this new descriptor, we conduct the face recognition by borrowing the idea of Universal Background Modeling (UBM) [7], which is widely used for speaker identification. First, a Gaussian Mixture Model based UBM is constructed for all the X-Y patches from the human face images of different subjects. The purpose of UBM is to encode the universal prior knowledge of X-Y patches. For a specific subject, Maximum A Posterior (MAP) adaptation [5] is used for adapting the UBM to specific subject based on all the X-Y patches extracted from all the images of this specific subject. Finally, the ratio of the likelihood from the adapted UBM for a specific subject and likelihood from the original UBM is used as score output for each specific subject, and the classification is conducted by comparing all these scores.

2. X-Y PATCHES

Local patches have been widely adopted for appearance modeling in computer vision literature [11]. Also recently, systems based on local patches followed by Bag Of Words (BOW) modeling have achieved great success in object recognition [9]. However, a major

*This research was funded in part by the U.S. Government VACE program, and in part by the NSF Grant CCF 04-26627. The views and conclusions are those of the authors, not of the US Government or its Agencies.

limitation of conventional patch based algorithm is that the spatial information is not well utilized while the spatial cue often plays an important role for visual appearance modeling [19]. Intuitively, the visual appearance modeling is the modeling of joint appearance and shape information. Motivated by these observations, we present for object representation a new descriptor, called X-Y patches which concatenate the x,y coordinates into the appearance features for a certain local patch, that is,

$$A_{XY} = (A, p_x, p_y), \tag{1}$$

where p_x and p_y are the coordinates, and A is the appearance feature vector of the local patch. In this paper, we use the DCT transform of the normalized pixel intensities as appearance features for A. The normalization is transforming each local patch to be of zero mean and unit variance, and we adopt dense sample grid in the image plane to extract all the X-Y patches. An illustration of the X-Y patch extraction is displayed in Figure 1.

3. UNIVERSAL BACKGROUND MODELING

After the feature extraction process, each face image is represented as an ensemble of X-Y patches. To model this ensemble in a vector space, Gaussian Mixture Model (GMM) [5] is a good solution owing to its flexibility and scalability. If the training set of each subject is sufficiently large, we can train a GMM for each subject. However, the training set of each subject is usually not able to cover all possible variations, such as different poses, illuminations and occlusions. To circumvent this issue of limited training set, we adopt the universal background modeling scheme which is widely used in speech/speaker recognition literature [7]:

$$\begin{aligned} P(x|\lambda) &= \sum_{i=1}^{M} w_i P_i(x|\lambda_i) && (2)\\ &= \sum_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i) && (3)\\ &= \sum_{i=1}^{M} w_i \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}||x-\mu_i||^2_{\Sigma_i^{-1}}} && (4) \end{aligned}$$

where $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_M]$ are the parameters of GMM, x is the extracted D-dimensional feature vector, e.g. A_{XY}, w_i is the prior probability of ith component, and $N(x|\mu_i, \Sigma_i)$ is a multivariate Gaussian density, with mean vector μ_i and covariance matrix Σ_i. Note that $\sum_{i=1}^{M} w_i = 1$. $P_i(\cdot|\lambda)$ denotes the likelihood function of the ith component in GMM. For simplicity, the covariance matrix Σ_i is usually set to be a diagonal matrix to lower the computation cost. The Maximum Likelihood (ML) estimation of the parameters can be obtained via Expectation Maximization (EM) algorithm [3].

Denote the log-likelihood of the sample x from GMM as

$$L(x|\lambda) = \log \sum_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i). \tag{5}$$

Let the posterior probability of each component given the observation z as $\gamma_i(x) = \frac{w_i N(x|\mu_i, \Sigma_i)}{\sum_{j=1}^{M} w_j N(x|\mu_j, \Sigma_j)}$, the log-likelihood of GMM for x can be rewritten as

$$\begin{aligned} L(x|\lambda) &= \log \sum_{i=1}^{M} w_i P_i(x|\lambda_i) && (6)\\ &= \log \sum_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i) && (7)\\ &= \sum_{j=1} \gamma_j(x) \log \sum_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i) && (8)\\ &= \sum_{j=1} \gamma_j(x) \log \frac{w_j N_j(x)}{w_j N_j(x)} \sum_{i=1}^{M} w_i N_i(x) && (9)\\ &= \sum_{j=1} \gamma_j(x) \log \frac{w_j N_j(x)}{\gamma_j(x)} && (10)\\ &= \sum_{j=1}^{M} \gamma_j(x) (\log N_j(x) + \log \frac{w_j}{\gamma_j(x)}) && (11)\\ &\leq \sum_{j=1}^{M} \gamma_j(x) \log N_j(x), && (12) \end{aligned}$$

where $N_i(x) = N(x|\mu_i, \Sigma_i)$ and (12) is from the Kullback-Leibler divergence property. From the derivation, we can see that the UBM can be roughly considered as a projection which projects x into the component space ($\log N_j(x)$ and computes the local scoring at each component ($\gamma_j(x)$).

4. FACE RECOGNITION BY UBM ADAPTATION

As mentioned in previous section, the UBM essentially defines a projection from original feature space to the component spaces. Two X-Y patches are considered similar only if the projected components are similar. This characteristic provides a natural mechanism for handling occlusion since the occluded X-Y patches are very different from the normal X-Y patches and consequently will not provide evidence for any subject in final classification.

The UBM is trained on a large amount of X-Y patches from different face images of different subjects. It essentially describes the distribution of general X-Y patches of human face. To generate subject-specific face model, we adopt the Maximum A Posterior (MAP) adaptation [5] as follows:

$$\begin{aligned} \gamma(i|A_{XY}, \lambda) &= \frac{w_i P_i(A_{XY}|\lambda)}{\sum_{j=1}^{M} w_j P_j(A_{XY}|\lambda)} && (13)\\ \gamma(i|\lambda) &= \sum_{p_x, p_y} \gamma(i|A_{XY}, \lambda) && (14)\\ \bar{\mu}_i &= \frac{1}{\gamma(i|\lambda)} \sum_{p_x, p_y} \gamma(i|A_{XY}, \lambda) A_{XY} && (15)\\ \hat{\mu}_i &= \mu_i + \frac{\gamma(i|\lambda)}{\gamma(i|\lambda) + \alpha} (\bar{\mu}_i - \mu_i) && (16) \end{aligned}$$

where $\gamma(i|A_{XY}, \lambda)$ is the posterior probability of the ith component given the observation A_{XY}. $\gamma(i|\lambda)$ is the soft count of observations which belong to the ith component. $\bar{\mu}_i$ is the sample mean of ith component given the training observations $\{A_{XY}\}_{t=1}^{T}$ and $\hat{\mu}_i$ is the adapted mean of ith component from the background mean μ_i. The smoothing factor $\frac{\gamma(i|\lambda)}{\gamma(i|\lambda)+\alpha}$ is designed to incorporate the number of observations into the final adapted mean. With this smoothing

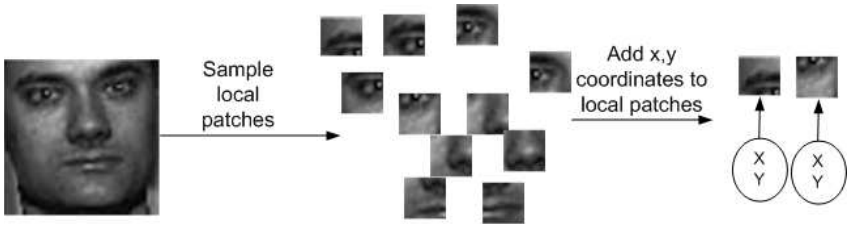

Fig. 1. An illustration of extracting the X-Y patches from one face image in CMU PIE [14] database.

Algorithm (%)	Train 5	Train 10	Train 15	Train 20
Eigenface	75.66	63.54	54.8	47.16
Fisherface	48.73	33.58	26.78	21.11
Laplacianface	62.68	36.15	23.76	17.65
LSDA	53.4	30.51	21.83	15.13
Patches	35.69	21.70	17.44	14.80
X-Y Patches	**31.13**	**16.11**	**11.3**	**9.24**

Table 1. Comparison (error rate) of face recognition algorithms on the CMU PIE database with different experiment configurations. Note that *Patches* means our algorithm based on patches without coordinate information.

factor, the adapted mean will adaptively adjust the mean vector according to the amount of observations on every component. If there are sufficient observations on one component, the adapted mean will rely more upon the sample mean for better data fidelity, otherwise it will more depend on the background mean. In testing stage, log-likelihood ratio between target subject and background models is used to score the test image, and the identity is determined as the subject with the largest score.

5. EXPERIMENTS

We evaluate the performance of the proposed X-Y patches based algorithm with real face recognition experiments. The state-of-the-art algorithms are mainly subspace learning based algorithms, such as Eigenfaces [15], Fisherfaces [1], Laplacianfaces [6], and LSDA [2]. We compare the proposed algorithm with these four most popular and latest subspace algorithms. Here, PCA is unsupervised while the other four are all supervised. We use the benchmark databases CMU PIE [14] for the experiments. We use Nearest Neighbor (NN) classifier for classification as conventionally.

5.1. PIE database

The CMU PIE (Pose, Illumination, and Expression) database contains in total 41368 images of 68 subjects with 500+ images for each. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination, and expression. For each subject, we manually select 168 images which cover large illumination variation, pose of roll/yaw/tilt head rotation and moderate variety in expression, constituting a challenging face database for recognition task. Face images are manually aligned, cropped out from the selected images and resized to be 20×20, with 256 gray levels per pixel. Figure 2 shows the sample images of two subjects from PIE database.

We use a subset of PIE containing five near frontal poses (C05, C07, C09, C27, C29) and all the images under different illuminations and expressions. In total, there are 170 images for each subject. To further reduce the size of the database, we randomly choose $1/5$ samples for each individual and obtain a subset with 34 images per individual. We finally have 2312 images in total. A random database partition is done with 5, 10, 15, and 20 images per subject for training, and the rest of the database for testing. Different subspace learning algorithms, PCA, LDA, LPP and LSDA, are applied to extract features for NN classification. For comparison, we also implement our algorithm based on patches without the coordinate information.

Table 1 shows the recognition results which are the average of 10 runs. From the results, we can have a set of interesting observations:

1. Our algorithm based on X-Y patches outperforms all the subspace learning algorithms in all the four experiment configurations, and achieves the lowest error rates of 31.13%, 16.11%, 11.3%, and 9.24% respectively.
2. Coordinate information is important for final classification, and X-Y patches based algorithm performs much better than the version without coordinate information.
3. Patch-based algorithm has the potential to outperform holistic subspace learning algorithm. Even without coordinate information, our algorithm still performs better than the other four subspace learning algorithms.

In order to test the performance of proposed method under partial occlusion, we randomly block some portion of the facial images from 0% to 60% of the total image area. The performance of proposed methods are reported in Table 2. Clearly, the results show that our proposed algorithm is robust against partial occlusion. Even under 60% occlusion, our proposed algorithm is still able to achieve comparable performance as the best subspace algorithm without occlusion.

6. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we proposed a new algorithm for face recognition. The major contributions are two-fold: a joint local appearance and shape

Fig. 2. Sample images of CMU PIE database.

Occlusion	0%	20%	40%	60%
20-Train	**9.24**	10.29	12.5	15.75
15-Train	**11.3**	13.23	17.95	20.66
10-Train	**16.11**	19.42	23.65	27.87
5-Train	**31.13**	34.18	39.04	41.37

Table 2. Face recognition performance under different percentage of occlusion for our X-Y patches based algorithm.

representation, namely X-Y patches, and a general learning and inferring framework based on the universal background modeling. The proposed X-Y patches based algorithm allows comparison between patches from different locations, which alleviates the requirement of exact pixel-wise alignment. The universal background modeling, on the other hand, provides a natural solution for handling partial occlusion. Combining these two major advantages, the proposed methods significantly outperform the state-of-art subspace learning algorithms.

Currently, we are planning to further verify the proposed algorithm on more databases, such as FERET [13] and FRGC [12]. Moreover, we are planning to conduct more experiments to evaluate the advantages of algorithm in: 1) robustness to the variation of pose, illumination, and expression, 2) combining multiple images for classification, 3) open-set experiment configuration, and 4) robustness to image misalignment issue.

7. REFERENCES

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on PAMI*, vol. 19, pp. 711-720, 1997.

[2] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," *Proc. Int. Joint Conf. on Artificial Intelligence*, 2007.

[3] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, pp. 1-38, 1977.

[4] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America*, vol. 14, pp. 1724-1733, 1997.

[5] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on SAP*, 1994.

[6] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 2000.

[8] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," *Proceedings of IEEE International Conference on Computer Vision*, 2001.

[9] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[10] C. Liu and H. Wechsler, "Comparative assessment of independent component analysis (ICA) for face recognition," *Proceedings of the Second International Conference on Audio- and Video-based Biometric Person Authentication*, pp. 211-216, 1999.

[11] S. Lucey and T. Chen, "Learning patch dependencies for improved pose mismatched face verification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 909-915, 2006.

[12] P. Phillips P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 947-954, 2005.

[13] P. Phillips, P. Rauss, and S. De, Feret (face recognition technology) recognition algorithm development and test report. *ARL-TR-995, U.S. Army Research Laboratory*, 1996.

[14] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illuminlation, and expression database," *IEEE Trans. on PAMI*, vol. 25, pp. 1615–1618, 2003.

[15] M. Turk and A. Pentland, "Face recognition using eigenfaces," *IEEE Conference on Computer Vision and Pattern Recognition*, 1991.

[16] L. Wiskott, J. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775-779, 1997.

[17] Y. Fu, M. Liu, and T. S. Huang, "Conformal embedding analysis with local graph modeling on the unit hypersphere,"*IEEE Conf. on Computer Vision and Pattern Recognition–The 1st Workshop on Component Analysis*, 2007.

[18] Y. Fu, J. Yuan, Z. Li, T. S. Huang, and Y. Wu, "Query-driven locally adaptive Fisher faces and expert-model for face recognition," *IEEE International Conference on Image Processing (IEEE ICIP'07)*, pp. 141-144, 2007.

[19] G-D. Guo and C. Dyer, "Patch-based image correlation with rapid filtering," *The 2nd Beyond Patches Workshop, in conjunction with the IEEE Conf. on CVPR'07*, 2007.