Texton Correlation for Recognition

Thomas Leung

Fujifilm Software
1740 Technology Drive, Suite 490, San Jose, CA 95110, U.S.A.
tleung@fujifilmsoft.com

Abstract. We study the problem of object, in particular face, recognition under varying imaging conditions. Objects are represented using local characteristic features called textons. Appearance variations due to changing conditions are encoded by the correlations between the textons. We propose two solutions to model these correlations. The first one assumes locational independence. We call it the conditional texton distribution model. The second captures the second order variations across locations using Fisher linear discriminant analysis. We call it the Fisher texton model. Our two models are effective in the problem of face recognition from a single image across a wide range of illuminations, poses, and time.

1 Introduction

Recognition under varying imaging conditions is a very important, yet challenging problem. Imaging conditions can change due to external and internal factors. External factors include illumination conditions (back-lit vs. front-lit or overcast vs. direct sunlight) and camera poses (frontal view vs. side view). Internal variations can arise from time (natural material weathers or rusts, or people aging) or internal states (facial expressions or a landscape changing appearance according to the season). The changes an object exhibits under varying imaging conditions are usually referred to as within-class variations in pattern recognition.

The ability to be invariant to within-class variations determines how successful an algorithm will be in practical applications. In recent years, a lot of attention in the research community has been devoted to this problem. Some representative examples and their application domains are (1) generic 3-D objects [11]; (2) faces [1,2,3,6,7]; and (3) natural materials [4,10,14].

In this paper, we strive to develop algorithms to recognize objects, in particular faces, under varying imaging conditions. The fundamental observation comes from human vision. After seeing many objects under different conditions, humans build an implicit internal model of how objects change their appearance. Using this model, humans can *hallucinate* any object's appearance under novel conditions. For example, one can easily recognize a person from the side after seeing only a single frontal picture of this person. Or, one can still recognize a friend with ease after not seeing him for 10 years. Of course, recognition is not always perfect, especially under some unusual conditions, but the accuracy is significant.

We adopt a learning framework to build a model of how the appearance of objects change under different imaging conditions. We call it the texton correlation model.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 203–214, 2004.

Textons are a discrete set of representative local features for the objects. The basic idea is to encode efficiently how the textons transform when illumination, camera pose, etc... change. Taking into account these transformations, we can build a similarity measure between images which are insensitive to imaging conditions. Using the texton correlation model, our algorithms can recognize faces from a single image of a person under a wide range of illuminations and poses, and also after many years of aging.

The outline of this paper is as follows. The concept of textons is reviewed in Section 2. Assuming locational independence, we propose a solution to capture the within-class variations using the *conditional texton distribution*. Experimental results using the conditional texton distribution are presented in Section 4. In Section 5, we introduce the idea of *Fisher textons* to capture second-order correlations across both pixel locations and imaging conditions. Results using *Fisher textons* are also presented. Finally, we conclude and discuss future work in Section 6.

2 Textons

Julesz [9] first proposed to use the term texton to describe the putative units of preattentive human texture perception. Julesz's textons — orientation elements, crossings, and terminators — lack a precise definition for gray level images. Recently, the concept of textons has been re-invented and operationalized. Leung and Malik [10] define textons as learned co-occurences of outputs of linear oriented Gaussian derivative filters. Variations of this concept have been applied to the problem of 3D texture recognition [4,10, 14]. We adopt a similar definition of textons in this paper.

What textons encode is a discrete set of local characteristic features of a 3D surface in the image space. This discrete set is referred to as the vocabulary. Every location on the image is mapped to an element in this vocabulary. For example, if the 3D surface is a human face, one texton may encode the appearance of an eye, another the mouth corner. For natural materials such as concrete, the textons may encode the image characteristic of a bar, a ridge, or a shadow edge. The textons can be learned from a single class (e.g. John, or concrete), thus forming a class-specific vocabulary. It can also be learned from a collection of classes (e.g. {John, Mary, Peter, ...}, or {concrete, velvet, plaster, ...}), thus forming a universal vocabulary. One advantage of the discrete nature of the texton representation is the ability to characterize changes in the image due to variations in imaging conditions easily. For example, when a person changes from a smiling expression to a frown, the mouth corner may change from texton element I to element J. Or when the illumination moves from a frontal direction to an oblique angle, the element on a concrete surface may transform from texton element A to element B. The main focus of this paper is to study how to represent this texton element transformation to recognize objects and materials under varying imaging conditions.

In this paper, textons are computed in the following manner. First, the image is filtered with a filterbank of linear Gaussian derivative filters. Specifically, the filters are the horizontal and vertical derivatives of circular symmetric Gaussian filters:

$$F_V(\sigma) = \frac{d}{dx} G_\sigma(x, y)$$

$$F_H(\sigma) = \frac{d}{dy} G_\sigma(x, y)$$
$$G_\sigma(x, y) = \frac{1}{2\pi\sigma} \exp(-\frac{x^2 + y^2}{2\sigma})$$

4 different scales are used, giving a total of 8 filters. The particular choice of filters is not very important[1]. This set is selected for their simplicity and ease of computation. In fact, the filters are x-y separable and filtering can be done in $O(N)$ time instead of $O(N^2)$ time, where N is the dimension of the kernel. After filtering, each pixel is transformed into a vector of filter responses of length 8. These filter responses are clustered using the K-means algorithm [5] to produce K prototypical features for the objects. With this vocabulary, every pixel in an image is mapped to the closest texton element according to the Euclidean distance in the filter output space. We call the output of this process the texton labels for an image. The value at each pixel is now between 1 and K, depending on which texton best describes the local surface characteristics.

3 Conditional Texton Distribution

We represent the texton transformation in a probabilistic formulation. The objective here is to capture how objects, faces, or natural materials change their appearance under varying imaging conditions. The goal is to learn the intrinsic transformation which is valid for all the instances within an object class. For example, in the context of face recognition, we would learn the transformation from a training set of a large group of people. The intrinsic variations within a single person and the differences between individuals are captured in the model. This learned transformation can be applied to any group of novel subjects and recognition can be achieved from a single image.

Let M be the image of a model. For example, in face recognition, M will be an image of the person you want to recognize. Let I be an incoming image. The task is to determine whether I is the same object as M. Let T_M be the texton labels for M and T_I be that of I. We define $P_{same}(T_I|T_M)$ to be the probability that I is the same object as the model M. Similarly, we define $P_{diff}(T_I|T_M)$ to be the probability that it is a different object. The likelihood ratio can be used to determine whether they come from the same object:

$$L(T_I|T_M) = \frac{P_{same}(T_I|T_M)}{P_{diff}(T_I|T_M)} \tag{1}$$

The task is to define $P_{same}(T_I|T_M)$ and $P_{diff}(T_I|T_M)$. We make the simple assumption that the texton labels are independent of their location:

$$P_{same}(T_I|T_M) = \prod_x P_{same}(T_I(x)|T_M(x))$$
$$P_{diff}(T_I|T_M) = \prod_x P_{diff}(T_I(x)|T_M(x))$$

[1] Other filter choices can be found in [10,13,14]

The likelihood ratio can be used as a similarity measure between an image and the model. We can either set a threshold on $L(T_I|T_M)$ to determine whether the face matches the model, or as in classification, assign the incoming image to the class with the highest likelihood ratio score, L.

3.1 Learning the Distribution from Data

The discrete nature of textons allows us to represent the distributions $P(T_I(x)|T_M(x))$ exactly, without making simplifying assumptions such as a Gaussian distribution. Notice that $T_I(x)$ is an element of the texton vocabulary and is scalar-valued: $T_I(x) \in \{1, \ldots, K\}$. In fact, with a texton vocabulary of size K, $P(T_I(x)|T_M(x))$ can be represented completely as an $K \times K$ table. This conditional probability table can be easily learned through training data.

Let the training set be $\mathcal{T}$. Let $\mathcal{C}_M$ be the set of all training data that belong to the same class as M. Let $a, b \in \{1, \ldots, K\}$ be two texton elements in the vocabulary. The entries in the probability table can be accumulated as follows[2]:

$$P_{\textit{same}}(T_I = a|T_M = b) = \frac{1}{Z_1} \sum_{M,I \in \mathcal{T}} \mathbf{1}_{(a,b,\mathcal{C}_M)}(T_I, T_M, I)$$

$$P_{\textit{diff}}(T_I = a|T_M = b) = \frac{1}{Z_2} \sum_{M,I \in \mathcal{T}} \overline{\mathbf{1}}_{(a,b,\mathcal{C}_M)}(T_I, T_M, I)$$

where Z_1 and Z_2 are normalizing constants to make $P_{\textit{same}}$ and $P_{\textit{diff}}$ probabilities. The function $\mathbf{1}_{(a,b,\mathcal{C}_M)}(T_I, T_M, C_I) = 1$ if $T_I = a, T_M = b, I \in \mathcal{C}_M$ and 0 otherwise. $\overline{\mathbf{1}}_{(a,b,\mathcal{C}_M)} = 1$ if $T_I = a, T_M = b, I \notin \mathcal{C}_M$ and 0 otherwise.

Applying these two learned conditional probability tables to the likelihood ratio $L(T_I|T_M)$ in Eq. 1, the similarity between a model and an incoming image can be computed.

4 Experiments

In this section, we will describe results of applying the conditional texton distribution model to the problem of face recognition. Before images are compared, faces are automatically extracted by a face detector [8]. Eyes and mouth corners are found using a similar algorithm to normalize each face for size and in-plane rotation. Each face is resampled to a standard size of 30×30 pixels. In all the experiments in this paper, separate texton vocabularies are learned independently at each pixel location. The main reason for this choice is the speed needed to compute texton assignment. For each pixel, a vocabulary size of 10 is used. In total, there are $30 \times 30 \times 10 = 9000$ textons altogether.

In all the experiments, the training set is used to obtain the texton vocabularies and the conditional texton distributions. All results are reported on a disjoint test set, in which none of the individuals appear in the training set. Results will be reported on two applications: face verification and face classification. First, we describe the databases used in this paper.

[2] The x dependency is implied implicitly to make the notations more readable.

4.1 Database

There are 3 face databases used to evaluate the performance under different imaging condition:

Yale face database: Using a geodesic dome with 64 flashes, researchers at Yale University collected a database of objects and human faces [2]. The set of face images is used in this paper to evaluate the performance of our algorithm towards varying illuminations. There are 15 individuals. Images in which the illumination angle is within $\pm 60^o$ in elevation and azimuth are used. These 32 images are shown in Figure 1. This database will be referred to as *yale* in this paper.

Fig. 1. The Yale face database. 32 illumination directions are used. The top two rows correspond to illumination directions randomly chosen to form the training set. The bottom two rows correspond to those in the test set.

FERET: The FERET training database [12] consists of faces of 1203 individuals, each with several frontal images. For a subset of the individuals (670 people), non-frontal images are available. We break up this database into two sets to measure the performance of our algorithm separately. The *Frontal FERET* consists of faces with different expressions and slightly different natural lighting. Robustness of our algorithm towards out-of-plane rotations (up to $\pm 45^o$) will be measured using the *Rotated FERET* subset.

ID photos: This database consists of employee ID photos at a Japanese company. There are 836 individuals, with a total of 1834 images. All the employees are Asians, with the majority Japanese. Every individual has at least 2 images, one taken at the time of hire and another taken recently[3]. For each individual, there is a large age difference in the different photographs — usually several years, up to as many as 20 years. This is a challenging database because people can change their appearance significantly over the span of several years: from wearing glasses to wearing contact lenses, from skinny to fleshy, from having a lot of hair to bald, etc. This database will be able to test how our algorithm performs when people change their appearance when aging. Since these are ID photographs, lighting is well-controlled, though not constant

[3] Some have one more photo taken during their employment.

across all photographs. Expression is usually neutral. This database will be referred to as *ID* in this paper. Due to privacy issues, only photos of 2 individuals can be shown in Figure 2.

Fig. 2. Examples of *ID* photos. Two different photographs of the same individual can be taken up to 20 years apart. This database can test our algorithm's performance against large age differences.

4.2 Face Verification

We first consider the problem of face verification. One application of face verification is an access control security system. A user inputs his/her identity through an ID card. A camera will capture an image of the user. The face will be detected and matched against the one previously stored in the system. If the match is good, the user will be granted access, otherwise denied. The basic concept is to compare two faces and decide whether it is the same person. In all our experiments, we randomly select two faces from the test set. These two faces can come from two different individuals or from the same individual[4]. Accuracy is measured by the false positive and false negative rates, in the form of a ROC (receiver operating characteristics) curve. All experiments are repeated using random partitions of the databases into training and test sets. Results reported are the average performance.

We first study the performance of our algorithm with respect to illumination variations using the *yale* database. 10 subjects are used for training and the remaining 5 subjects for testing. 16 illumination directions are chosen randomly and the corresponding images are used as training. These illumination directions are the same for every training individual. The corresponding images for one person are shown in the top two rows in Figure 1. The bottom two rows show the illumination directions for the test set. There is a complete disjoint between the training set and the test set. In other words, *none* of the illumination directions in the test set is present in the training set for any individual. Our algorithm performs perfectly in this experiment, getting 0% false positive and 0% false negative. Notice that the two images to be verified can have light directions up to 120^o apart. This means the texton distribution does a very good job encoding the intrinsic changes in feature appearance under different illuminations. The learned distribution extrapolates well for both new individuals and novel illumination directions.

[4] For the case of the same individual, the two images are never identical.

The performance of our algorithm on the *ID*, *Frontal FERET*, and *Rotated FERET* databases is shown in Figure 3. Figure 3(b) is the zoomed-in version of (a). The blue, red, and green curves are the ROC curves for the *Frontal FERET*, *Rotated FERET*, and *ID* databases respectively. The equal error rate (false positive rate equals to false negative rate) for the three experiments are 4.2%, 10.5%, and 9% respectively.

The dashed black curve in Figure 3 is the ROC curve on the *Frontal FERET* database with a system trained using *ID* pictures. The idea is to see how well the learned texton distributions generalize to a different database. Most machine learning algorithms guarantee good generalization only if the statistics of the training and test sets is identical. In this case, the statistics can be quite different, for example, the ethnic composition of the subjects are totally different: *ID* contains pictures of predominantly Japanese, while *Frontal FERET* contains pictures with diverse ethnic backgrounds.

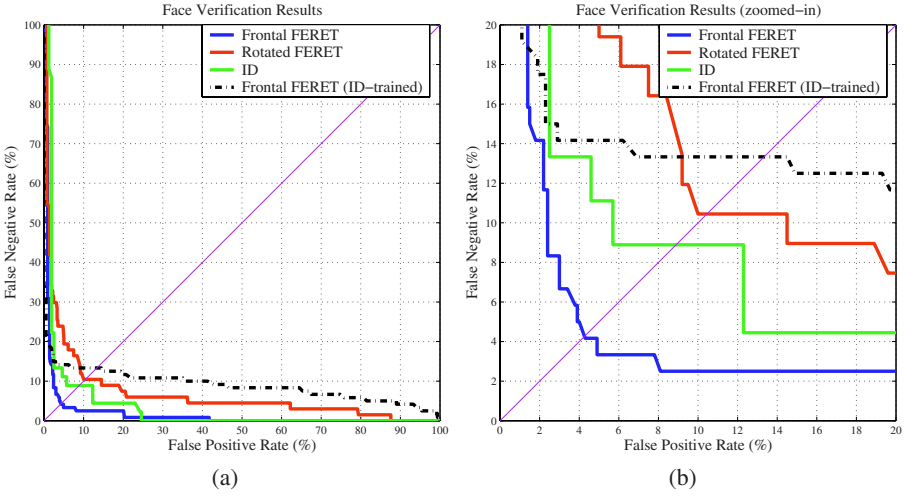

Fig. 3. Face verification results. The blue, red, and green curves are the ROC curves for the *Frontal FERET*, *Rotated FERET*, and *ID* databases respectively. For each of these three experiments, the training and test sets come from the same database. The dashed black curve indicates the ROC curve for the *Frontal FERET* database using a system trained with the *ID* photos.

4.3 Face Classification

In this section, we investigate the effectiveness of our conditional texton distribution model for the problem of face classification. Let there be P individuals, each with a single image as the model. For any new image of these P people, we want to automatically determine who he is. We will use the similarity measure in Section 3 (Equation 1) and classify this new image into the model with the highest similarity score.

For all the databases, we randomly pick P individuals from the test set. For each person, one image is randomly selected to be the model, another to be the probe. The

model and the probe can be of vastly different imaging conditions. This procedure is repeated multiple times for different choices of P individuals and different random training and test sets. Our algorithm works perfectly for the *yale* database, giving 0% error. The results for the other databases are reported in Figure 4. The blue, red, and green curves report the error rates for the *Frontal FERET*, *Rotated FERET*, and *ID* databases respectively. In these three cases, the training and test sets come from the same database. We would like to emphasize that for all the experiments, only a single image is used for the model. This is a very difficult problem, especially for the *Rotated FERET* case, because the probe can be up to 90^o out-of-plane rotated from the model. The black curve in Figure 4 presents the error rate on *Frontal FERET* with a system trained using the ID photos. This indicates the effectiveness of our algorithm when the statistics of the test set (predominantly Japanese) is different from that of the training set (ethnically diverse).

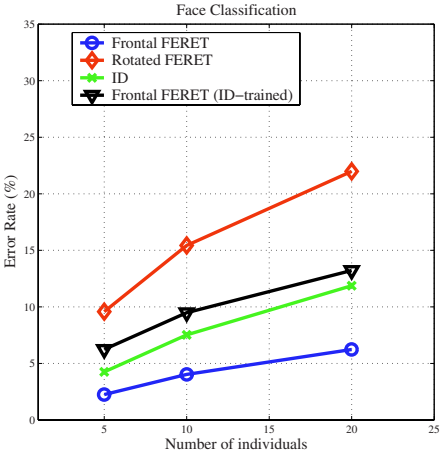

Fig. 4. Face classification results. The model consists of one image. The blue, red, and green curves represent the error rates for the *Frontal FERET*, *Rotated FERET*, and *ID* databases respectively. The black curve represents the error rate for the *Frontal FERET* with a system trained using *ID* photos.

5 Fisher Textons

The conditional texton distribution model presented in Section 3 makes the assumption that the texton assignments are independent of location. This is obviously a wrong assumption. For example, the appearance of the left eye and the right eye are definitely correlated. However, this assumption enables us to compute the likelihood ratio (Equation 1) efficiently.

In this section, we explore the correlation between locations on the face. Specifically, we take into account second order correlations. After texton assignment, every face is

turned into a 30×30 array of texton labels. Each pixel, $T_I(x)$, takes on the value between $1, \ldots, K$, where K is the size of the texton vocabulary at each pixel[5]. We transform each pixel to an indicator vector of length K: $[0, \ldots, 0, 1, 0, \ldots, 0]$ with 1 at the k^{th} element if $T_I(x) = k$. We concatenate all the vectors together, so that each image becomes a $30 \times 30 \times 10 = 9000$ dimensional vector.

We perform the Fisher linear discriminant analysis [5] on these vectors to obtain the projection directions which are best for separating faces from different people. Specifically, let the within-class scatter matrix be:

$$\begin{aligned} \boldsymbol{S_w} &= \sum_{i-1}^{c} \boldsymbol{S_i} \\ \boldsymbol{S_i} &= \sum_{\boldsymbol{x} \in \mathcal{C}_i} (\boldsymbol{x} - \boldsymbol{m_i})(\boldsymbol{x} - \boldsymbol{m_i})^t \\ \boldsymbol{m_i} &= \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{C}_i} \boldsymbol{x} \end{aligned}$$

where c is the number of classes and $\mathcal{C}_i$ is the set of training examples belonging to class i. $n_i = |\mathcal{C}_i|$ and $n = \sum_i^c n_i$. The between-class scatter matrix is defined as:

$$\boldsymbol{S_b} = \sum_{i=1}^{c} n_i (\boldsymbol{m_i} - \boldsymbol{m})(\boldsymbol{m_i} - \boldsymbol{m})^t$$

where $\boldsymbol{m}$ is the total mean vector:

$$\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{c} n_i \boldsymbol{m_i}$$

The objective is to find $\boldsymbol{V}$ to maximize the following criterion function:

$$J(\boldsymbol{V}) = \frac{|\boldsymbol{V}^t \boldsymbol{S_b} \boldsymbol{V}|}{|\boldsymbol{V}^t \boldsymbol{S_w} \boldsymbol{V}|}$$

The columns of the optimal $\boldsymbol{V}$ are the generalized eigenvectors corresponding to the largest eigenvalues in

$$\boldsymbol{S_b} \boldsymbol{v_i} = \lambda_i \boldsymbol{S_w} \boldsymbol{v_i}$$

The vectors $\boldsymbol{v_i}$ are the projection directions which capture the essential information to classify objects among different classes. The idea is that when a large number of training examples are used, the $\boldsymbol{v_i}$'s can distinguish between people not only those present in the training set.

We call these projection vectors $\boldsymbol{v_i}$ the *Fisher textons*. Every incoming image is transformed into an N dimensional vector by projecting into these Fisher textons. Similarity between two faces is taken simply as the Euclidean distance between the projections.

[5] $K = 10$ in this paper.

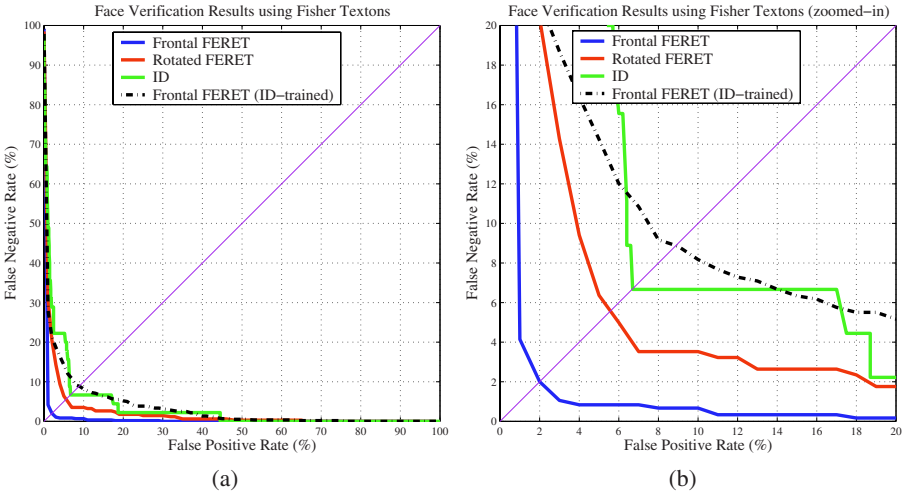

Fig. 5. Face verification results using Fisher textons. The blue, red, and green curves are the ROC curves for the *Frontal FERET*, *Rotated FERET*, and *ID* databases respectively. For each of these three experiments, the training and test sets come from the same database. The dashed black curve indicates the ROC curve for the *Frontal FERET* database using a system trained with the *ID* photos.

Let us contrast the differences between the Fisher textons and the conditional texton distribution. For Fisher textons, locational correlations are captured up to second order. However, imaging condition correlations are captured only up to the second order as well. On the other hand, the texton conditional distribution model sacrifices location dependencies to capture the exact texton distributions under changing imaging conditions.

The results on the problem of face verification are shown in Figures 5. The blue, red, and green curves are the ROC curves for the *Frontal FERET*, *Rotated FERET*, and *ID* databases respectively. The equal error rates are 2%, 5.5%, 6.6% respectively. The dashed black curve indicates the ROC curve for the *Frontal FERET* database using a system trained with the *ID* photos, with an equal error rate of 9%. The performance on the face verification task is uniformly better than that produced by the conditional texton distribution model. The added locational dependencies more than offset the sacrifice made on the imaging condition correlations.

Results for face classification using Fisher textons are shown in Figure 6. The performance is better for the *Frontal FERET* (blue curve) and *Rotated FERET* (red curve) databases. However, it is worse for *ID* (green curve) and *Frontal FERET* database when trained using *ID* photos (black curve). One possible explanation is that the within-class variations in the *ID* database is so large (because of the long timeline) that just capturing the second order correlations is not enough to distinguish individuals well. Using the whole distribution, as in the case of the conditional texton distribution model, can thus produce better results.

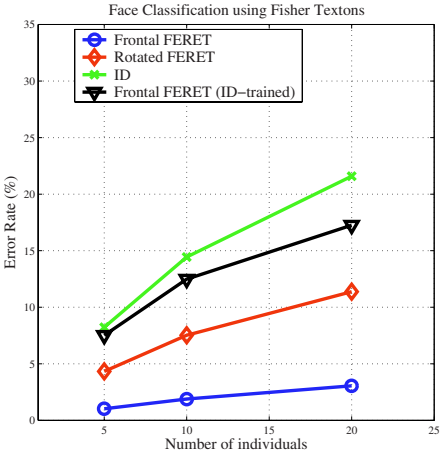

Fig. 6. Face classification results using Fisher Textons. The blue, red, and green curves represent the error rates for the *Frontal FERET*, *Rotated FERET*, and *ID* databases respectively. The black curve represents the error rate for the *Frontal FERET* with a system trained using *ID* photos.

6 Conclusions

In this paper, we study the problem of recognition under varying imaging conditions. Two algorithms are proposed based on the idea of textons. The first algorithm is to use conditional texton distributions to model the within-class and between-class variations exactly. But, the assumption of locational independence is made. We call the second algorithm Fisher textons. Second order correlations in both location and imaging condition variations are captured. Both algorithms are effective in the problems of face verification and recognition.

Comparing with state-of-the-art algorithms is difficult without training and testing on the same datasets. Future work includes thorough comparisons with other algorithms. Another direction for future work is to develop algorithms to capture the exact dependencies from both pixel locations and changing imaging conditions. The obvious choice is a Markov Random Field. However, the parameters in MRFs are difficult to estimate without a large quantity of data. Inference is not a trivial task either. Finding efficient ways to capture the correlations will be an interesting problem from both theoretical and practical points of view.

Acknowledgements. Portions of the research in this paper use face images collected at Yale University and under the FERET program. The author would like to thank David Forsyth, Jitendra Malik, Sergey Ioffe, Troy Chinen, Yang Song, and Ken Brown for helpful discussions.

References

1. V. Blanz and T. Vetter. Face recognition basedo n fitting a 3d morphable model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9), 2003.
2. H. Chen, P. Belhumeur, and D. Jacobs. In search of illumination invariants. In *Proc Conf CVPR*, 2000.
3. T. Cootes, K. Walker, and C.J. Taylor. View-based active appearance models. In *Proc. Intl. Conf. Automatic Face and Gesture Recognition*, 2000.
4. O. Cula and K. Dana. Compact representation of bidirectional texture functions. In *Proc. Computer Vision and Pattern Recognition*, pages 1041–7, 2001.
5. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2001.
6. A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under varying lighting and pose. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6), 2001.
7. R. Gross, L. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. In *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, 2002.
8. S. Ioffe. Automatic red-eye reduction. In *Proc. Int. Conf. Image Processing*, 2003.
9. B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–7, March 1981.
10. T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimenional textons. *Int. J. Computer Vision*, 43(1):29–44, 2001.
11. H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal on Computer Vision*, 14(1):5–24, 1995.
12. P. Phillips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.
13. C. Schmid. Constructing models for content-based image retrieval. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2001.
14. M. Varma and A. Zisserman. Classifying images of materials: achieving viewpoint and illumination independence. In *Proc. European Conference Computer Vision*, pages 255–71, 2002.