Face Image Retrieval Using HMMs

Aleix Martínez

Robot Vision Lab - School of Electrical and Computer Engineering
Purdue University, IN 47906
and
Sony Computer Science Laboratory - 75005 Paris (France)
aleix@ecn.purdue.edu

Abstract

This paper introduces a new face recognition system that can be used to index (and thus retrieve) images and videos of a database of faces. New face recognition approaches are needed because, although much progress has been made to identify face taken from different viewpoints, we still cannot robustly identify faces under different illumination conditions, or when the facial expression changes, or when a part of the face is occluded on account of glasses or parts of clothing. When face recognition methods have worked in the past, it was only when all possible "image variations" were learned. Principal Components Analysis (PCA) and Fisher Discriminant Analysis (FDA) are well-known cases of such methods.

In this paper we present a different approach to the indexing of face images. Our approach is based on identifying frontal faces and it allows reasonable variability in facial expressions, illumination conditions, and occlusions caused by eye-wear or items of clothing such as scarves. We divide a face image into n different regions, analyze each region with PCA, and then use a Bayesian approach to finding the best possible global match between a query image and a database image. The relationships between the n parts is modeled by using Hidden Markov Models (HMMs).

1 Introduction

The rapidly increasing interest in Content Base Image Retrieval (CBIR) systems is fueled by the possibilities that are envisioned in many diverse domains. One such fertile domain consists of retrieving face images from large multimedia databases, as might be produced by television and newspaper establishments. In this paper, we present an approach that uses Hidden Markov Models (HMMs) to identify and index face images from a database of static and dynamic (video) images.

In the past, several researchers have worked with the Principal Components Analysis (PCA) [4, 15] based approaches for face identification. However, in general, the face images of a multimedia database may have been taken under different illumination conditions; might correspond to different facial expressions; may suffer from different degrees of occlusion (wearing scarf or sun-glasses is seen as an occlusion problem); and the images may have been taken from different viewpoint. Therefore, the identification procedure has to be designed in such a way as to be unaffected by all these distortions. Although much progress has been made in identifying face images taken from different viewpoints (e.g. [16, 7]), not a whole lot has yet been done to account for different illumination conditions [10, 1] and facial expressions [7, 1] and almost none on resolving the occlusion problem.

As a solution to the illumination problem, [10] proposed to eliminate the three largest eigenvectors of the eigenspace because it has been empirically shown that these three largest eigenvectors mainly represent changes of the illumination conditions. The main problems with this approach are: (a) a priori it is not known either there are or there are not illumination changes in the given sample images (note that if there are no changes the elimination of these three largest eigenvectors will make the system learn a much worse eigenspace because these three vectors will be the ones that best describe the sample faces); (b) it has not been shown that for *every* possible combination of illumination changes this approach is correct; and (c) we are still losing some information because not all the information in these three first eigenvectors is related to the illumination conditions. [1] and [14], took another route proposing the use of Fisher Discriminant Analysis (FDA) instead. The main advantage of FDA versus PCA is that it allows us to find those features of the space that are most discriminant, rather than those that are most descriptive (in a linear sense). However, even then, the illumination problem can only be *resolved* by adding samples of all possible illumination variations (notice that a priori we might not know or

Figure 1. (a) The sample image used for training. (c) The new incoming image to be identified. (b,d) A local part of the images.

might not be able to predict all possibilities).

The other problems that arise in face identification – problems caused by variations in facial expressions and problems caused by occlusion – can only be solved by adding samples with such distortions into the database used for training. But, unfortunately, the logistics of collecting data sometimes make it too difficult to include samples corresponding to all possible variations in the training set. Clearly, assuming that figure 1(a) was used for training, figure 1(b) could hardly be identified by means of any of the above techniques (when the whole mug-face is used). A better approach consists of using local areas, as we will explain in this contribution. Our approach divides all pre-localized faces (the localization step is done manually) into n different parts (in this paper we make $n = 6$) and project the raw information of each of these local parts into the eigenspace (PCA-space). [7] has also proposed the use of a local PCA, but only a voting strategy is used to find the best global match. Our approach makes use of a Gaussian mixture model in a HMM approach to better overcome the above defined problems.

Our overall approach presented in this paper consist of assuming that there is an underlying identity associated with the observed sequence, and then invoke a HMM to find that model (identity) that best fits the given observation. Although it is true that up to this point we could also use a Markov Model (MM) instead of a HMM to compute the given observations, we will see that given the fact that faces can appear under different illumination conditions and partial occlusions, the observable sequence cannot directly match with a given model, but that several possibilities have to be tested (which means that the identity reminds hidden, and that HMM are needed).

HMMs have proved to be very successful in speech recognition (see for example [11]). HMMs have also been used in computer vision, as for example for learning the saccadic movements produced by an active vision system [12]; for American Sign Language (ASL) recognition [13]; for learning and recognizing facial expressions [8]; and for modeling human behaviors [9].

In the rest of this paper, Section 2 presents the local PCA feature extraction procedure; Section 3 derives the HMM procedure used for learning and recognition (the mixture of Gaussian approach is also described); Section 4 reports the experimental results obtained to day; and, finally, conclusions are in section 5.

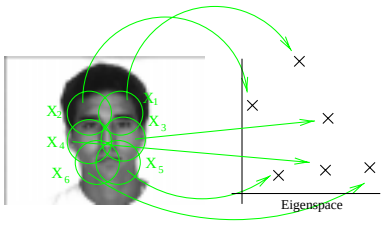


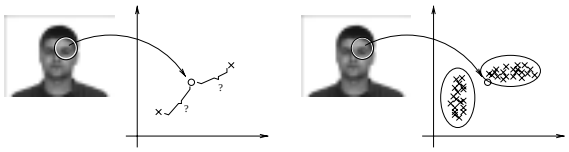
Figure 2. The face is divided into six different local areas. Each area is projected into a global eigenspace.

2 Local feature extraction using a density model approach

As shown in figure 1, it seems that it would be more appropriate to match local features than the features that depend on the whole image (the global features). We therefore propose to divide face images into n different parts, analyze them independently (locally), and finally merge all this information using a HMM approach. In this paper we make $n = 6$. Figure 2 shows an example of this local feature extraction process.

Each of these local areas are projected into the eigenspace where learning takes place. To build the eigenspace, we first learn it using all the local areas obtained from all the images in the database. Each area is described by a vector ($\mathbf{x}_i$) and then normalized such that $\|\mathbf{x}_i\| = 1$. The average of all vectors ($\mathbf{c}$) is also subtracted from each vector $\mathbf{x}_i$ to ensure that the eigenvector associated with the largest eigenvalue represents the dimension in the eigenspace in which variance of images is maximized (in a correlation sense). The vector $\mathbf{X} = \{\mathbf{x}_1 - \mathbf{c}, ..., \mathbf{x}_r - \mathbf{c}\}$ is used to define the covariance matrix: $\mathbf{Q} = \mathbf{X}\mathbf{X}^T$. The eigenspace is obtained by taking the m eigenvectors associated with the largest eigenvalues of $\lambda_i \mathbf{e}_i = \mathbf{Q}\mathbf{e}_i$. These m eigenvectors define the projection matrix $M = \{\mathbf{e}_1, ..., \mathbf{e}_m\}$.

One problem that we have to contend with is that face localization (a step that is carried out manually at the time an image is entered into the database) cannot be precise enough to guarantee that the extracted local information will always be projected into the correct area of the eigenspace (corresponding to the correct sample). Figure 3(a) shows an example of that. In order to avoid this problem, not only do we extract information from the selected local area, but also from its adjacent pixels (using a $p \times q$-size window). As shown in figure 3(b) this creates a distribution that can be modeled using a mixture of Gaussians learned using the EM algorithm [2].



(a) (b)

Figure 3. (a) We cannot know which of the two points is the best match. (b) Using a PDF approach, we can give a probability value in a more robust manner.

3 Using HMM to learn and recognize faces

As pointed out in the introduction, it is worthwhile to assume that there is an underlying identity that although not directly observable can nonetheless be observed using a (or a set of) stochastic process. We use the previously defined eigenspace as the observation process.

3.1 Hidden Markov Modeling

A time domain process demonstrates a Markov property if the conditional probability of the current event, given all present and past events, depends only on the jth most recent events. When the current event depends solely on the most recent past event, we say that the process is a first order Markov process, i.e. $P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...) = P(q_t = S_j | q_{t-1} = S_i)$, where S_i denotes the ith state (of a finite state machine) and q_t the state value at time t (i.e. the actual state). This probability, denoted in the sequel as a_{ij}, corresponds to the probability of being in state S_i at time $t-1$ and at state S_j at time t. Typical restrictions in a Markov chain are: $a_{ij} \geq 0$ and $\sum_{j=1}^{N} a_{ij} = 1$ (where N is the total number of states). We can formally define an HMM by its probabilities measures $\lambda(A, B, \pi)$, its states $S = \{S_1, \ldots, S_N\}$ and its associated symbols (the alphabet) $V = \{v_1, ..., v_M\}$ (where $A = \{a_{ij}\}$ defines the transition probabilities, $B = \{b_j(k)\}$ defines the probability of having symbol v_k while in state S_j, and $\pi_i = P(q_1 = S_i)$ defines the initial state probabilities).

However, we do not have a discrete alphabet V, but a continuous observation density instead. The most general PDF used in a HMM is the mixture of Gaussian [11]. In such a case, an HMM is essentially a mixture model encoding information about the history of a time series in the value of a multinomial variable – the hidden state. Then, we redefine:

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm} \mathcal{G}(\mathbf{O}, \mu_{jm}, \mathbf{\Sigma}_{jm}) \qquad (1)$$

where $\mathbf{O}$ is the vector being modeled, c_{jm} is the mth mixture coefficient in state j, and $\mathcal{G}$ is the multivariate Gaussian density function (μ_{jm} and $\mathbf{\Sigma}_{jm}$ are its associated mean and covariance matrix). M is the total number of models used in each state (this number is assumed to be known and in consequence must be given to the system). The c_{jm} coefficients satisfy the restrictions $\sum_{m=1}^{M} c_{jm} = 1$ and $c_{jm} \geq 0$, so that the PDF is properly normalized, i.e. $\int_{-\infty}^{\infty} b_j(\mathbf{O}) d\mathbf{x} = 1$.

It has been shown [5] that the re-estimation formulas (based on the EM algorithm [2]) for c_{jm}, μ_{jm} and $\mathbf{\Sigma}_{jm}$ are of the form:

$$c_{jm}^{[z+1]} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(j,k)}$$

$$\mu_{jk}^{[z+1]} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot \mathbf{O}_t}{\sum_{t=1}^{T} \gamma_t(j,k)}$$

$$\mathbf{\Sigma}_{jk}^{[z+1]} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (\mathbf{O}_t - \mu_{jk}^{[z]})(\mathbf{O}_t - \mu_{jk}^{[z]})^T}{\sum_{t=1}^{T} \gamma_t(j,k)}$$

where $\gamma_t(j,k)$ is the probability of being in state j at time t with the kth mixture component accounting for $\mathbf{O}_t$, which can be written as:

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \right] \left[\frac{c_{jk}\mathcal{G}(\mathbf{O}_t, \mu_{jk}, \mathbf{Q}_{jk})}{\sum_{m=1}^{M} c_{jm}\mathcal{G}(\mathbf{O}_t, \mu_{jm}, \mathbf{Q}_{jm})} \right]$$

where $\alpha_t(j)$ is the forwards variable and $\beta_t(j)$ the backwards variable of the HMM used to reestimate the probability measure previously defined. The forwards variable can be iteratively defined as: $\alpha_1(j) = \pi_j b_j(O_1)$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1})$$

and the backwards variable as: $\beta_T(i) = 1$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j).$$

3.2 Learning

The above method can be used to learn the sample face images. When doing so, the re-estimation formulas for the HMM parameters ($\bar{\pi}_i$ and $\bar{a}_{ij}$) can be written as: $\bar{\pi}_i = \gamma_1(i)$ and

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} (\alpha_t(i) a_{ij} b_j(\mathbf{O}_{t+1}) \beta_{t+1}(j)) / P(\mathbf{O}|\lambda)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

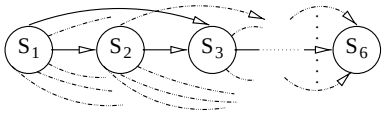


Figure 4. A HMM consists of six states each belonging to the six local parts of a face. All forward transitions might be used, but this will depend on the training data set.

where $P(\mathbf{O}|\lambda) = \sum_{i=1}^{N}\sum_{j=1}^{M}\alpha_t(i)a_{ij}b_j(\mathbf{O}_{t+1})\beta_{t+1}(j)$ and T the number of observations.

For each image of each individual, the system creates an HMM model (see figure 4). The states of each HMM correspond to each of the six local parts. The use of a HMM instead of a simple MM is justify because observations need not contain the six local parts, but can only have some of theme (the others part being occluded or not present). Thus, the observation sequence is not known and the identity of each observation remains hidden and can only be observed by probabilistic models.

This process ends up with $P * H$ HMMs, where P is the number of people in the database and H the number of sample images per person. However, a more desirable case would correspond on having a single HMM per person. In order to achieve that all HMMs of the same person are merged together (in a single one). In doing so, the new transition probabilities (between models) are set to $1/H$. This also implies that HMMs have to be used; since many combinations are allowed, the underlying identity remains hidden.

In the future, it might be of interest to use an active learning procedure, such as e.g. [12], to *on-line* update these transition probabilities as new images are introduced into the database.

3.3 Indexing and Recognition

To index a new image for entry into the database, we first need to identify the class to which it belongs. This recognition process consists on finding the HMM that best represents a given observation sequence. The normal way to achieve this is by computing the probability of a new observation sequence $\mathbf{O}$ given a possible model λ, i.e. $P(\mathbf{O}|\lambda)$. The model λ that maximizes this probability infers the (hidden) class value. The usual way to find the best matching is using the Viterbi algorithm [11].

3.4 Recognition in video sequences

The above recognition process can be improved when recognizing faces in a video sequence. Assuming that all face images of the sequence are frontal faces, we can use a longer observation $\mathbf{O}$ (where the observation is now a sequence of images under slightly different viewing/environmental conditions). When using this approach, the above definition of models has to be extended by linking the last state (of each learned HMM) to the first state (of each learned HMM).

4 Experimental results

In order to test the accuracy of the indexing system described, a new face database was built. It includes face images of over 100 individuals [6]. Faces (as shown in figure 5) appear at different facial expressions, illumination conditions and occlusions (wearing sun-glasses or a scarf). Images with occlusions are also under different illumination conditions. All images were taken by the same camera under tightly controlled conditions of illumination and viewpoint (all images correspond to frontal faces). All images are of 768 by 576 pixels and of 24 bits depth (RGB color) stored in "raw" files (this database is publicly available from http://rvl1.ecn.purdue.edu/~aleix/). For this research, we also processed 30 video sequences. Each sequence consists of 25 images and almost all of them containing a frontal face (only those images with frontal faces were used in the experiments reported in this paper – this can be readily achieved using a frontal face detector). When running the tests, 50 people (25 males and 25 females) were randomly selected from the database. Images were converted to gray-level images (by adding all three color channels) and sampled at half their size.

Five different tests were run: (i) learning using all images (a-l) and recognizing by "erasing" one of the local features, i.e. making one of the local areas equal to random noise with mean zero (all images are tested with all possible "erasing", i.e. six possibilities), (ii) learning with images (a-f) and recognizing images (g-l), (iii) learning with images (g-l) and recognizing with images (a-f), (iv) learning with (a,f), recognizing with (b-e,g-l), and (v) learning with images (a-f) recognizing using all 30 sequences.

All tests were done using the six states HMM described above. Each state containing a single Gaussian model. The PCA space was of six dimensions. $p = 4$ and $q = 4$.

Test #	Lear.	Rec.	Rec. rate
i	a-l	a-l + noise	96.83%
ii	a-f	g-l	98.5%
iii	g-l	a-f	97.1%
iv	a,f	b-e,g-l	72%
v	a-f	sequences	28/30 (93.5%)

5 Conclusions

In this paper, we have introduced an indexing approach based on the identification of frontal face images. Faces are allowed to appear at different illumination conditions, facial expressions and occlusions. A

Figure 5. Our face database. (a) smile. (b) Sad. (c) Screaming. Illuminations: (d) right, (e) left, (f) both. Wearing glasses with illumination conditions: (g) normal, (h) right, (i) left. Wearing scarf with illumination conditions: (g) normal, (h) right, (i) left.

Bayesian approach was used to find the best match between a sequence of "local" observations and the learned "local" features models. The use of HMMs has allowed the system to achieve recognition even when the conditions under which the new incoming image was taken did not correspond to the conditions previously encountered during the learning phase. This last point is very important, because it gives a new "generalization" capability to the system.

Acknowledgments

Thanks to Prof. Avi Kak for stimulating discussion. I am also very grateful to Prof. Leah Jamieson and Mike Johnson for allowing and helping me to use the HTK software.

References

[1] P.N. Belhumeour, J.P. Hespanha and D.J. Kriegman. *Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection.* IEEE Tr. **PAMI**-19(7):711-720, 97.

[2] A.P. Dempster, N.M. Laird and D.B. Rubin. *Maximum likelihood from incomplete data via the em algorithm.* Journal Royal Statistical S., 77.

[3] K. Fukunaga. *Introduction to statistical pattern recognition (second edition).* Academic Press, 90.

[4] M. Kirby and L. Sirovich. *Application of the Karhunen-Love procedure for the characterization of human faces.* IEEE Trans. **PAMI**-12(1):103-108, 90.

[5] L.A. Liporace. *Maximum likelihood estimation for multivariate observations of Markov sources.* IEEE Trans. **IT**-28(5):729-734, 82.

[6] A. Martínez and R. Benavente. *The AR face database.* CVC Tach.Rep. #24, June 98.

[7] B. Moghaddam and A. Pentland. *Probabilistic visual learning for object representation.* IEEE Trans. **PAMI**-19(7):696-710, 97.

[8] N. Oliver, A. Pentland and F. Bérard. *LAFTER: Lips and face real time tracker with facial expression recognition.* Proc. CVPR, 97.

[9] N. Oliver, B. Rosario and A. Pentland. *A Bayesian computer vision system for modeling human interactions.* Proc. ICVS, 99.

[10] A. Pentland, T. Starner, N. Etcoff, N. Masoiu, O. Oliyide and M. Turk. *Experiments with eigenfaces.* looking at people, workshop of IJCAI 93.

[11] L.R. Rabiner. *A tutorial on hidden Markov models and selected applications in speech recognition.* Proc. IEEE 77(1):257-285, 89.

[12] R.D. Rimey and C.M. Brown. *Controlling eye movements with hidden Markov models.* Int. J. Comp. Vision 7(1):47-65, 91.

[13] T. Starner and A. Pentland. *Visual recognition of american sign language using hidden Markov models.* Int. workshop AFGR, 95.

[14] D.L. Swets and J.J. Weng. *Using discriminant eigenfeatures for image retrieval.* IEEE Trans. **PAMI**-18, No. 8:831-836, 96.

[15] M. Turk and A. Pentland. *Eigenfaces for recognition.* Journal Cognitive Neuroscience, 91.

[16] T. Vert. *Synthesis of novel views from a single face image.* Int. J. Comp. Vision, 98.