Clustering Face Images with Application to Image Retrieval in Large Databases

Florent Perronnin and Jean-Luc Dugelay

Institut Eurécom
Multimedia Communications Department
2229 route des Crêtes, BP 193
06904 Sophia-Antipolis Cédex, FRANCE

ABSTRACT

In this article, we evaluate the effectiveness of a pre-classification scheme for the fast retrieval of faces in a large image database. The studied approach is based on a partitioning of the face space through a clustering of face images. Mainly two issues are discussed. How to perform clustering with a non-trivial probabilistic measure of similarity between faces? How to assign face images to all clusters probabilistically to form a robust characterization vector? It is shown experimentally on the FERET face database that, with this simple approach, the cost of a search can be reduced by a factor 6 or 7 with no significant degradation of the performance.

Keywords: Biometrics, Face Recognition, Indexing, Clustering.

1. INTRODUCTION

Defining a meaningful measure of similarity between face images for the problem of automatic person identification and verification is a very challenging issue. Indeed, faces of different persons share global shape characteristics, while face images of the same person are subject to considerable variability, which might overwhelm the inter-person differences. Such variability is due to a long list of factors including facial expressions, illumination conditions, pose, presence or absence of eyeglasses and facial hair, occlusion and aging. A measure of similarity between face images should therefore be rich enough to accommodate for all these possible variabilities. Although using a more complex measure may improve the performance, it will also generally increase the computational cost. Hence, it is difficult to design a measure which is both *accurate* and *computationally efficient*. However, both properties are required to tackle the very challenging task of automatic retrieval of face images in large databases. Mainly, two techniques based on the notion of *coarse classification* have been suggested to reduce the number of comparisons when searching a database.

The first approach makes use of two (or even more) complementary measures of distance and *cascades* them. The first distance, which has a low accuracy but requires little computation, is run on the whole dataset and the N-best candidates are retained. The second distance, which has a high accuracy but requires more computation, is then applied on this subset of images. Such an approach has already been applied for instance to the problem of multimodal biometrics person authentication.[1]

The second approach consists in *partitioning the image space*, e.g. by *clustering* the dataset. When a new target image is added to the database, one computes the distance between this image and all clusters and the image is associated to its nearest cluster. When a query image is probed, the first step consists in determining the nearest cluster and the second step involves the computation of the distances between the query image and the target images assigned to the corresponding cluster. It is interesting to notice that the pre-classification of images is an issue which has received very little attention from the face recognition community. For other biometrics, such as fingerprints, this has been a very active research topic.[2]

The quality of a pre-classification scheme can be measured through the *penetration rate* and the *binning error rate*.[3] The penetration rate can be defined as the expected proportion of the template data to be searched under

Send correspondence to Professor Jean-Luc Dugelay: E-mail Jean-Luc.Dugelay@eurecom.fr, Telephone: +33 (0)4.93.00.26.41, Fax: +33 (0)4.93.00.26.27

the rule that the search proceeds through the entire partition, regardless of whether a match is found. A binning error occurs if the template and a subsequent sample from the same user are placed in different partitions and the binning error rate is the expected number of such errors. Both target and query images can be assigned to more than one cluster. Indeed if face images of a given person are close to the "boundary" between two or more clusters, as large variabilities may not be fully handled by the distance measure, different images of the same person may be assigned to different clusters as depicted on Figure 1. To solve this problem, target and query images can be assigned to their K nearest clusters or to all the clusters whose distance falls below a predefined threshold. Obviously, the decrease of the binning error rate is obtained at the expense of an increase in the penetration rate.

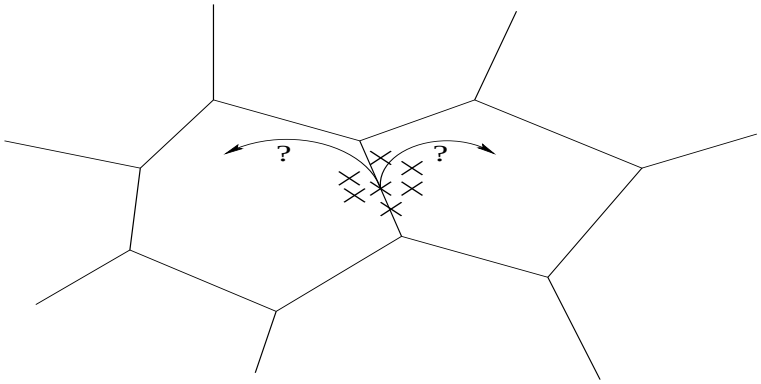


Figure 1. Uncertainty in cluster assignment.

The main contribution of this paper is to evaluate the reduction of the amount of computation which can be achieved when searching a face database using a pre-classification strategy based on clustering. In this work, the measure of similarity between face images that is considered is the Probabilistic Mapping with Local Transformations (PMLT) introduced in.[4] This approach consists in estimating the set $\mathcal{R}$ of possible transformations between face images of the same person. The global transformation is approximated with a set of local transformations under the constraint that neighboring transformations must be consistent with each other. Local transformations and neighboring constraints are embedded within the probabilistic framework of a 2-D HMM. The states of the HMM are the local transformations, emission probabilities model the cost of a local mapping and transition probabilities the cost of coherence constraints (c.f.[4,5] for more details). The measure of similarity between a template image I_t and a query image I_q is $P(I_q|I_t, \mathcal{R})$, i.e. the likelihood that I_q was generated from I_t knowing the set $\mathcal{R}$ of possible transformations. This approach was shown to be robust to facial expression, pose and illumination variations.[5] Even if its computational complexity is low enough to perform real-time verification (which is a one-to-one matching) or even identification (which is a one-to-many matching) for a target set that does not exceed a few hundred of images on a modern PC, it is still too high for searching large face databases.

Therefore, we will first have to consider the issue of clustering face images with this non-trivial probabilistic measure of similarity. Many clustering algorithms, especially those based on a probabilistic framework, can be directly interpreted as an application of the Expectation-Maximization (EM) algorithm.[6] During the E-step, the distance between each observation and each cluster centroid is computed and each observation is assigned to its nearest cluster (or probabilistically to all clusters). During the M-step, the cluster centroid is updated using the assigned observations. The update step also depends on the chosen distance since the centroid is defined as the point that minimizes the average distance between the assigned observations and the centroid. When using simple metrics the update step is greatly simplified. For instance, for the Euclidean distance, the update step is a simple averaging of the assigned observations. In the case of complex distances, such as PMLT, computing the centroid is much more challenging.

Then, we will consider the possibility to assign each face image to all clusters probabilistically instead of assigning face images in a hard manner. A similar approach, referred to as *anchor modeling* has already been

proposed in the field of automatic speaker detection and indexing.[7] Another contribution of this paper is to improve over the original anchor modeling approach.

The remainder of this paper is organized as follows. In section 2, we describe the face image clustering procedure. In section 3, we briefly review the anchor modeling approach and propose two improvements. In section 4, we present experimental results before drawing conclusions in section 5.

2. CLUSTERING FACE IMAGES

As our measure of similarity is probabilistic, it is natural to use a Maximum-Likelihood (ML) framework to perform clustering.[8] In this section, we first briefly present the basic ML approach based of the EM principle. We then discuss the issue of cluster initialization and propose a fast procedure which is tailored to the problem of interest.

2.1. EM-based Clustering

Our goal is, given a set of N images $\{I_1, ..., I_N\}$, to estimate C clusters $\{\mathcal{C}_1, ..., \mathcal{C}_C\}$. The measure of distance between an image I_n and a cluster is the probability that this image was generated by the cluster centroid knowing $\mathcal{R}$. Therefore *images* $\{I_1, ..., I_N\}$ *are naturally treated as query images* and *cluster centroids as templates*. In the following, we will denote by O_n the "query representation" of I_n, i.e. the set of feature vectors extracted from I_n. O will denote the set of all observations: $O = \{O_1, ..., O_N\}$. We will denote by λ_c the "template representation" of the centroid of cluster $\mathcal{C}_c$. Thus, the measure of distance between I_n and cluster $\mathcal{C}_c$ is $P(O_n|\lambda_c, \lambda_\mathcal{R})$.

We assume that the distribution of the data can be modeled with a mixture of C components, where each component corresponds to one of the C clusters:

$$P(O_n|\lambda, \lambda_\mathcal{R}) = \sum_{c=1}^{C} w_c P(O_n|\lambda_c, \lambda_\mathcal{R}) \tag{1}$$

with $\lambda = \{w_1, ..., w_C, \lambda_1, ..., \lambda_C\}$. The mixture weights w_c are subject to the following constraint:

$$\sum_{c=1}^{C} w_c = 1 \tag{2}$$

We also assume that samples are drawn independently from the previous mixture.

$$P(O|\lambda) = \prod_{n=1}^{N} P(O_n|\lambda) \tag{3}$$

Our goal is to find the parameters $\{w_1, ..., w_C\}$ and $\{\lambda_1, ..., \lambda_C\}$ which maximize $P(O|\lambda)$. This problem cannot be solved directly and an iterative procedure based on the EM algorithm is generally used. The application of the EM algorithm to the problem of the estimation of mixture densities is based on the computation (E-step) and maximization (M-step) of Baum's auxiliary $\mathcal{Q}$ function.[6] The hidden variable includes both the state sequence Q, i.e. the set of local transformations which are "chosen" when measuring the similarity between a image and a cluster centroid (c.f. section 1), and a variable Θ that indicates the mixture component (i.e. the cluster assignment). Therefore, the $\mathcal{Q}$ function takes the following form:

$$\mathcal{Q}(\lambda|\lambda') = \sum_{Q} \sum_{\Theta} P(Q, \Theta|O, \lambda') \log P(O, Q, \Theta|\lambda) \tag{4}$$

where λ' is the current parameters estimate and λ is the improved set of parameters that we seek to estimate. If we split $\log P(O, Q, \Theta|\lambda)$ into $\log P(O, Q|\Theta, \lambda) + \log P(\Theta|\lambda)$, the $\mathcal{Q}$ function can be written as:

$$\mathcal{Q}(\lambda|\lambda') = \sum_{c=1}^{C} \sum_{n=1}^{N} \gamma_n^c \log(w_c) + \sum_{c=1}^{C} \sum_{n=1}^{N} \gamma_n^c \sum_{Q} \log P(O_n, Q|\lambda_c, \lambda_\mathcal{R}) \tag{5}$$

where the probability γ_n^c for image I_n to be assigned to cluster $\mathcal{C}_c$ is given by:

$$\gamma_n^c = P(\lambda'_c|O_n, \lambda_{\mathcal{R}}) = \frac{w'_c P(O_n|\lambda'_c, \lambda_{\mathcal{R}})}{\sum_{i=1}^{C} w'_i P(O_n|\lambda'_i, \lambda_{\mathcal{R}})} \tag{6}$$

To maximize $\mathcal{Q}(\lambda|\lambda')$, we can maximize independently the two terms. To find the optimal estimate $\hat{w}_c$ of w_c, we maximize the first term under the constraint (7) and obtain:

$$\hat{w}_c = \frac{1}{N} \sum_{n=1}^{N} \gamma_n^c \tag{7}$$

The maximization of the second term does not raise technical difficulties. However, as this issue is not the focus of this paper, the details are not presented here and the interested reader can refer to.[5]

2.2. A Fast Initialization Procedure

While the EM procedure is bound to reach a local optimum, it is by no means guaranteed to reach the global one. The quality of the optimum which is found depends on several factors, one of which is the initialization of cluster centroids. Indeed, after preliminary experiments, it was clear that selecting the initial centroids in a random manner could lead to very different solutions.

A simple procedure we employed to alleviate this problem was to perform as an initialization step a *hierarchical agglomerative* clustering.[8] The goal is not to obtain the C best possible clusters but to obtain with a fast procedure reasonable seed centroids that can be subsequently fed to the EM procedure described in the previous section. The basic idea of agglomerative clustering is to start with N clusters, each cluster containing one image, and to merge the clusters until the desired number of clusters C is obtained.

Therefore, a distance between clusters needs to be defined. Let $\{I_n\}$ be a set of images assigned to $\mathcal{C}_i$. The likelihood $\mathcal{L}(\mathcal{C}_i)$ of $\mathcal{C}_i$ is given by:

$$\mathcal{L}(\mathcal{C}_i) = \sum_{n:I_n \in \mathcal{C}_i} P(O_n|\lambda_i, \lambda_{\mathcal{R}}) \tag{8}$$

As we want a fast initialization procedure, we do not want to have to use the EM procedure to estimate λ_i. Thus we make use of the concept of *medoid*[9]: one chooses the most likely observation among the set of observations assigned to $\mathcal{C}_i$. Thus, if λ_{I_m} is the "template representation" of I_m, then:

$$\lambda_i = \arg \max_{m:I_m \in \mathcal{C}_i} \sum_{n:I_n \in \mathcal{C}_i} P(O_n|\lambda_{I_m}, \lambda_{\mathcal{R}}) \tag{9}$$

Let us remind that, during the initialization step, the goal is to find the C cluster centroids which maximize the likelihood of the set of observations. After each merging stage, the likelihood of the set of observations will decrease. Therefore, our goal is to merge at each step of the agglomerative clustering the two clusters that lead to the *smallest decrease* of the likelihood. Hence, the distance between two clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ is defined as the decrease in likelihood after the merging:

$$\mathcal{D}_{like}(\mathcal{C}_i, \mathcal{C}_j) = \mathcal{L}(\mathcal{C}_i) + \mathcal{L}(\mathcal{C}_j) - \mathcal{L}(\mathcal{C}_i \cup \mathcal{C}_j) \tag{10}$$

Note that this is similar to the criterion which is often used by Gaussian merging algorithms.[10] While at each step we are guaranteed to obtain the smallest decrease in likelihood, we are not guaranteed that the sequence of steps leads to the global maximum.

However we found experimentally that if we apply directly this procedure, the clusters we obtain may be highly unbalanced, i.e. some clusters may be assigned a large number of data items while others may contain only a small number of data items. This is a problem as a cluster centroid cannot be robustly estimated with a too small number of data items. Hence, we should penalize the previous distance in order to take into account

the balance between clusters. Let n_i be the number of data items in cluster $\mathcal{C}_i$ and let N be the total number of data items. We also introduce $p_i = n_i/N$. Clearly, the entropy[11]:

$$\mathcal{H} = -\sum_{i=1}^{N} p_i \log(p_i) \tag{11}$$

is a measure of balance as, the larger $\mathcal{H}$, the more balanced is the set of clusters. Let $\mathcal{H}$ be the entropy for the set of clusters $\{\mathcal{C}_1, ...\mathcal{C}_C\}$. If we merge clusters $\mathcal{C}_i$ and $\mathcal{C}_j$, then the delta entropy will be:

$$\Delta\mathcal{H}(\mathcal{C}_i, \mathcal{C}_j) = p_i \log(p_i) + p_j \log(p_j) - (p_i + p_j) \log(p_i + p_j) \tag{12}$$

which is a negative quantity. The closer is this quantity to zero, the smaller the reduction of entropy, and thus the smaller the reduction of the "balance" in our system.

Hence, we use as a measure of distance between two clusters $\mathcal{C}_i$ and $\mathcal{C}_j$:

$$\mathcal{D}(\mathcal{C}_i, \mathcal{C}_j) = \mathcal{D}_{like}(\mathcal{C}_i, \mathcal{C}_j) - \rho\Delta\mathcal{H}(\mathcal{C}_i, \mathcal{C}_j) \tag{13}$$

where ρ is a positive parameter that keeps the balance between the two possibly competing criteria: the minimum likelihood decrease versus the maximum entropy decrease.

3. PROBABILISTIC ASSIGNMENT OF FACE IMAGES

In this section, we first briefly review the anchor modeling approach. We then suggest two significant improvements over the original approach.

3.1. A Brief Review of Anchors for Indexing

A limitation of the multiple cluster assignment paradigm described in the introductory section is that it does not make the most out of the available information. To make our argument clear, let us assume that the face space is partitioned as depicted on Figure 2. When the template image I_t is added to the database, it is likely to be assigned to clusters $\mathcal{C}_6$, $\mathcal{C}_7$ and $\mathcal{C}_8$. At test time, the query image I_q is first assigned to $\mathcal{C}_2$, C_8 and C_9 and then compared to all the template images contained in one of these clusters, which includes I_t. However, I_t and I_q are fairly distant and, thus, unlikely to belong to the same person. Therefore, such a comparison will most likely be wasteful. The reason why I_t and I_q were compared while they should not have been is that, when assigning an image to one or multiple clusters, we throw away a lot of valuable information: the "distances" $p(I|\mathcal{C}_n)$. Indeed, the vector $v = [p(I|\mathcal{C}_1), ...p(I|\mathcal{C}_N)]^T$ could be used to characterize a face image I.

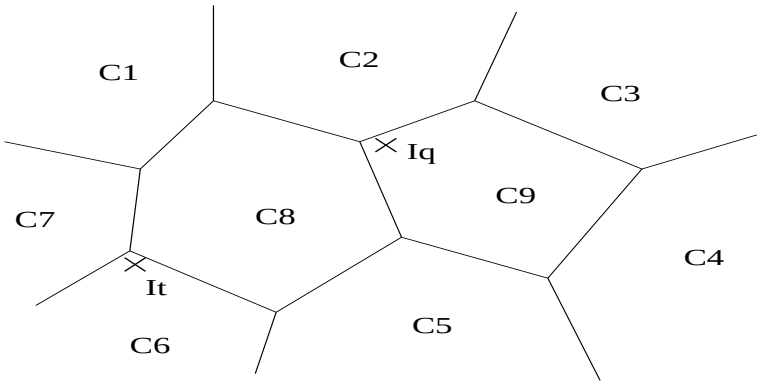


Figure 2. Case where a template image I_t and a query image I_q are unlikely to belong to the same person but are still assigned to the same cluster.

Anchor modeling was proposed in the fields of speaker detection and indexing[7]: a speech utterance s is scored against a set of models $\{A_1, ..., A_N\}$ referred to as *anchors* and the vector $v = [p(s|A_1), ..., p(s|A_N)]^T$ is used to

characterize the speech utterance. This characterization vector can be understood as a projection of the target image into a speaker space. Let v_q be the characterization vectors of I_q. Then at test time, we first compute the distance between v_q and the characterization vectors of all template images contained in the database. Although there are as many distances to compute as template images, this is very fast as these vectors are low dimensional. Then I_q is compared with the template images I_t that are less than a given threshold distant from I_q. Note that this approach can be seen as a special case of the cascading approach. Indeed, characterization vectors are simplified representations of face images and thus recognition based purely on these vectors has a low accuracy. However, they are fairly fast to estimate and very fast to compare. An interesting property of such a cascading approach is that the characterization vector retains the properties of the costly distance, a property that is not discussed in.[7] Indeed, if the distance is robust to some variations, then the characterization vector should not be significantly affected by these variations.

3.2. Improving the Original Anchor Approach

We propose in this paper two significant improvements over the original anchor modeling approach:

- As in[7] the number of anchor models was large (668 in their experiments), methods for reducing the size of the Euclidean distance comparison were investigated in an effort to increase performance by using only those anchor models that provide good characterizing information. However, such an approach does not reduce the cost of computing v which can also be significant. In the proposed approach, our anchors are not faces but the centroids which are obtained after clustering a set of face images. The clustering step should therefore perform a dimension reduction and drastically decrease the cost of computing v and of comparing it with other vectors.
- Instead of using a characterization vector v based on the likelihood, we propose to use posterior probabilities: $v = [p(\mathcal{C}_1|I), ..., p(\mathcal{C}_N|I)]^T$. Such a vector should be more robust, especially to a mismatch between training and test conditions, as it normalizes the likelihood.

4. EXPERIMENTAL RESULTS

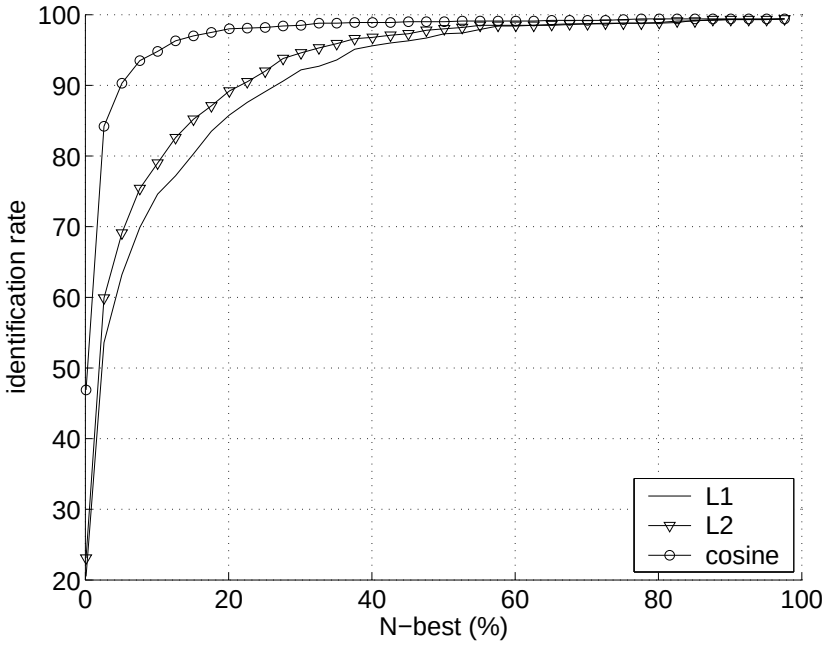
In this section, we first describe the experimental setup. We then compare the performance of posterior-based characterization vectors with likelihood-based vectors. Finally, we evaluate the impact of the reduction of the number of anchors on the efficiency of the retrieval.

4.1. Experimental setup

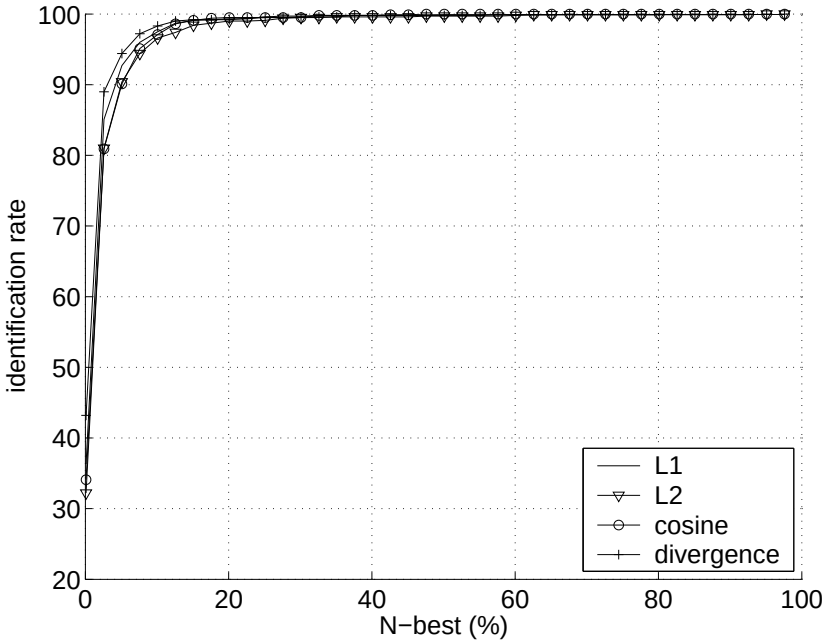
Our experiments were carried out on the FERET face database, a standard testbed for face recognition algorithms. The choice of this database was primarily motivated by the high number of individuals it contains (almost 1,200). 695 persons were chosen randomly for the training and 500 person for the test phase. We use for all training and test persons 2 images, those that are labeled FA and FB and which correspond to frontal images with variations in facial expression. The measure of similarity based on PMLT was trained exactly as described in[4] with the 1,390 training images. The estimation of cluster centroids was performed on the same data. At test time, each of the 1,000 face images was chosen successively as the query and the 999 remaining images were used as templates. The baseline performance of our system is the identification rate when each query is compared to all the templates, which is 95.7%. On a 2 GHz Pentium 4 with 1 GB RAM, this set of comparisons takes on the order of 5 seconds. The goal is now to reach a similar performance but with a number of comparisons which is significantly smaller than 999.

4.2. Evaluating the Impact of Posterior-based Characterization Vectors

If we want the pre-classification step to be effective, the distance between characterization vectors should be based on relatively simple metrics. The goal of the first set of experiments is to determine 1) which distance is the most appropriate to measure the similarity of characterization vectors and 2) whether the characterization vector based on posteriors is superior to the one based on likelihoods. Thus, in this first set of experiments, we perform identification with the characterization vectors only. We tested the L_1 (city-block), L_2 (Euclidean) and



(a)



(b)

Figure 3. Performance of a system with $C = 20$ clusters which makes use of (a) log-likelihood-based characterization vectors (b) posterior-based characterization vectors. Cumulative identification rate versus N-best (as a percentage of the database).

.

cosine metrics on both types of characterization vectors. As a posterior-based characterization vector defines a discrete probability distribution, we also tried the symmetric divergence on this type of vectors.

Note that the likelihoods $P(O_n|\lambda_c, \lambda_{\mathcal{R}})$ are extremely large (on the order of $10^{10,000}$) and thus they are difficult to compare directly. Therefore, in the following we did not use likelihood-based characterization vectors but characterization vectors based on the log-likelihood. In the same manner, $P(O_n|\lambda_c, \lambda_{\mathcal{R}})$'s are so large that the posteriors $P(\lambda_c|O_n, \lambda_{\mathcal{R}})$ are equal to 1 for the most likely centroid and 0 for the other ones. Thus, to increase the fuzziness of the assignment, we raised the posteriors to the power of a small positive factor β and then renormalized them so that they would sum to unity. In the following experiments we set $\beta = 0.01$.

Results are presented for $C = 20$ clusters on Figure 3. On Figure 3 (a), we compare the performance of the L_1, L_2 and cosine metrics for characterization vectors based on the log-likelihood. Clearly, the cosine is by far the best choice. On Figure 3 (b), we compare the performance of the L_1, L_2, cosine and symmetric divergence metrics for posterior-based characterization vectors. Results are much improved for the first three metrics (especially for L_1 and L_2) compared to log-likelihood-based vectors. The four measures of distance exhibit a similar performance but the symmetric divergence seems to outperform the three other metrics by a slight margin. Hence, in the following experiments, we will use posterior-based characterization vectors and the similarity of two such vectors will be measured with the symmetric divergence.

4.3. Evaluating the Impact of a Reduction of the Number of Anchors

Now that we have chosen the type of characterization vector and the metric, we can evaluate the performance of our system when characterization vectors are used during a pre-classification step to find the most likely candidates. We present results for various numbers of clusters as the identification rate versus the percentage of comparisons compared to the baseline case where we perform an exhaustive search (Figure 4). Note that we have to take into account the comparisons with the C cluster centroids and the comparison with all the templates that are retained after the pre-classification. While the increase of performance from 5 to 10 clusters is very significant, especially for a small number of comparisons, it is smaller when going from 10 to 20 clusters. No improvement could be obtained with more than 20 centroids. This shows that, for the problem of interest, clustering is very important as only a very small number of clusters is required for an efficient pre-classification. The best performance we could obtain was a reduction of the computational cost by a factor 6 or 7 with no significant degradation of the performance compared to an exhaustive search.

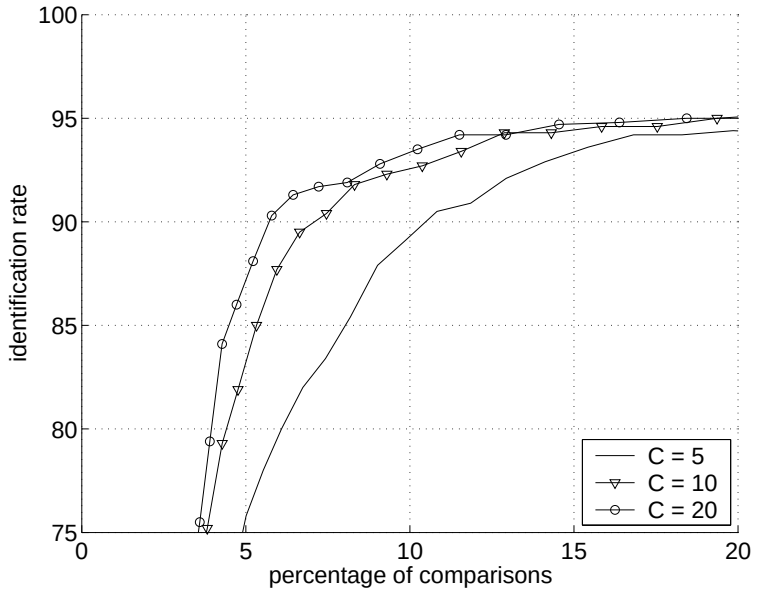


Figure 4. Performance of the system with probabilistic cluster assignment for a varying number C of clusters.

5. CONCLUSION

In this article, we evaluated the effectiveness of a pre-classification scheme for the fast retrieval of faces in a large database. We studied an approach based on a partitioning of the face space through a clustering of face images. We discussed mainly two issues. First, we addressed the problem of clustering face images with a non-trivial measure of similarity. As the chosen measure is probabilistic, we naturally used a ML framework based on the EM principle. Then, we discussed how to form a characterization vector, which could be used for an efficient indexing, by concatenating the distances between the considered image and all cluster centroids. While this is similar to anchor modeling, we suggested to significant improvements over the original approach. Experiments carried out on the FERET face database showed that, with this simple approach, the cost of a search could be reduced by a factor 6 or 7 with very little degradation of the performance.

Although the exact figures might vary depending on the specific database or measure of similarity, we believe that they give a reasonable idea of the speed-up which can be expected with a pre-classification approach. While this is a very significant cost reduction, it is clear that such a scheme would not be sufficient for databases which contain millions of faces. For such a challenging case, other approaches would have to be considered in combination with the studied approach. Especially, the use of multiple hardware units or of exogenous data (such as the gender or the age)[12] would most certainly be necessary.

ACKNOWLEDGMENTS

The authors would like to thank Professor Kenneth Rose from the University of California at Santa Barbara (UCSB) for drawing their attention to the important clustering issue. The authors would also like to thank France Telecom Research and Development for partially funding their research activities.

REFERENCES

1. L. Hong and A. Jain, "Integrating faces and fingerprints for person identification," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* **20**, pp. 1295–1307, Dec 1998.
2. A. Jain and S. Pankanti, *Advances in Fingerprint Technology*, ch. Automated Fingerprint Identification and Imaging Systems. CRC Press, 2nd ed., 2001.
3. A. Mansfield and J. Wayman, "Best practices in testing and reporting performance of biometric devices," Aug 2002.
4. F. Perronnin, J.-L. Dugelay, and K. Rose, "Deformable face mapping for person identification," in *IEEE Int. Conf. on Image Processing (ICIP)*, **1**, pp. 661–664, 2003.
5. F. Perronnin, *A Probabilistic Model of face Mapping Applied to Person Recognition*. PhD thesis, Intitut Eurécom, 2004.
6. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society* **39**(1), pp. 1–38, 1977.
7. D. Sturim, D. Reynolds, E. Singer, and J. Campbell, "Speaker indexing in large audio databases using anchor models," in *Proc. of the IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, **1**, pp. 429–432, 2001.
8. R. Duda, P. Hart, and D. Stork, *Pattern classification*, John Wiley & Sons, Inc., 2nd ed., 2000.
9. L. Kaufman and P. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, ch. Partitioning around medoids. John Wiley & Sons, 1990.
10. A. Sankar, "Experiments with a Gaussian merging-splitting algorithm for HMM training for speech recognition," in *Proc. of the 1997 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 99–104, 1998.
11. T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1993.
12. A. Jain, S. Pankanti, L. Hong, A. Ross, and J. Wayman, "Biometrics: a grand challenge," in *Proc. of the IEEE Int. Conf. on Pattern Recognition (ICPR)*, **2**, pp. 935–942, 2004.