ABSTRACT

We propose a system to “retrieve” the mental image of a face from a large database using Bayesian inference and relevance feedback. Since the “target image” exists only in the mind of the user, mental image retrieval differs sharply from standard, example-based retrieval and has not been widely studied. In designing the relevance feedback engine, we adopt probabilistic models for the display and answer processes. The answer model is designed to capture properties of human cognition in choosing among displayed faces. The images in each display are selected according to heuristics inspired by maximizing the conditional mutual information between the answer and the target given the previous feedback. Simulations and real tests validate show that the feedback engine operates in real-time and locates the target in a reasonable number of displays.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—retrieval models, relevance feedback

General Terms
Algorithms

1. INTRODUCTION

In traditional face retrieval problems [11, 17], a variation of the target image is available in some physical format and the search engine sorts the images in the face database according to their closeness to the example in hand. However, in many practical cases, there is no actual “example image” [3]. Instead, what users have are merely subjective opinions and impressions about the appearance of the target based on their latent memory view – a “mental face image.” We propose a system to interactively retrieve mental faces via the framework of relevance feedback. This scenario has extensive applications, e.g., to information retrieval, security, e-business and web-based browsing.

There are many documents on image retrieval with relevance feedback [16, 4, 10, 21, 8]. However, as described in [22], most of them involve category search with many samples for each category; moreover, in many cases, some low level features, such as texture, color and shape, are sufficient to distinguish among different categories. Hence, the problem of retrieval could be posed as incremental pattern classification through the process of relevance feedback. However, mental face retrieval is based on the identity of an individual face, a “target search” problem. Moreover, there are typically very few examples of each subject (category) in the database in comparison with the number of categories, which makes schemes developed for category search virtually impossible to apply.

In a benchmark work on mental image retrieval as target search, Cox et al. adopted a Bayesian model and information-theoretic tools [5]. Geman and Moquet [9] generalized this framework for a toy application involving searching for polygons. In both cases, only pairwise comparisons are considered. For large face databases, these schemes are unrealistic due to the complexity of mental face matching and facial features.

Little has been reported about mental face retrieval. Navaret et al. proposed a solution with self-organizing maps in [13] without providing statistical results. Masaomi et al. initialize their exploration with line-drawings of faces with around ten features and three variations per feature [14], which greatly simplifies the feature space. However, their system inevitably deteriorates with real face databases. In our work, a complete system is proposed based on a comprehensive analysis of relevance feedback in the context of mental face matching, accounting for both human psychology and efficient search. The performance is validated by experiments with human subjects.

In Section 2 we introduce our framework for mental face retrieval. Signature extraction and face metrics are described in Section 3. The relevance feedback model is given in Section 4 and experimental results in Section 5. In Section 6, we draw conclusions and propose future work.

2. SYSTEM FRAMEWORK

The system consists of two parts: off-line signature extraction and on-line interactive retrieval, as illustrated in Fig.2. The signatures of face images are extracted offline and form the feature database used for comparing images during retrieval. Usually, the opinions or impressions of users about face images are very high-level (e.g., of a semantic
is feasible to find mechanisms for comparing images which generally reflect both the criteria employed by typical users and the underlying numerical signatures and metrics used. Similarly, it is unreasonable to ask the user to provide low-level image signatures. Due to this "semantic gap", machine. For example, it is unrealistic to require the system to search for a sincere old man with big eyes and a flat nose. Similarly, it is unreasonable to ask the user to provide the subspace signature of his mental target. Nevertheless, it is feasible to find mechanisms for comparing images which generally reflect both the criteria employed by typical users and the underlying numerical signatures and metrics used by the system. In particular, the description "(displayed) person A looks like my target" can be translated into "facial image of A is close to the target in signature space". Consequently, the query consists of a set of images (referred to as the "display") and the user is expected to provide feedback on the perceived similarity between his mental image and those displayed. Hence, the selection of signatures and metrics is driven by trying to obtain high coherence between the user’s response and the metrics in signature space. Since face recognition is also a task about judgments on individual identity based on the appearance of facial images, we examined several kinds of signatures originally developed for face recognition.

Relevance feedback involves a sequence of iterations of the query/answer process. There is then an "answer model" for integrating the response of the user and a "display model" for choosing which images to display. In the query stage, driven by the display model algorithm, the system presents a question to the user about his target. In the answer stage, the model tells the machine how to interpret the answer of the user in terms of the metric on images in the database. The query/answer process is iterated until the system actually displays an image of the user’s target. Usually this requires considerable peaking of the distribution over images maintained by the system. Hence, the objective of the feedback machine is to minimize, subject to a fixed number of displayed images at each step, the number of iterations until the target is displayed. By applying relevance feedback in mental retrieval, knowledge about the target is obtained through the query/answer process rather than via an actual images as in standard query-by-example. The “semantic gap” is then bridged by sequential learning. More precise (but still nontechnical) descriptions of the system components now follow:

- **Answer Model:**
  The answer should provide relative information about how “close” the target is to the set of displayed images. The possible types of answers vary according to precision of the desired information. The answer can be as precise as the degree of similarity between each displayed image and the target; it can be a rough, qualitative evaluation (for example, “similar/dissimilar”) for each displayed image; or it can be a single, overall comparative response (for example, “A is closest to my target”). Two types of answer strategies were investigated: individual relevant/irrelevant and overall comparative responses. The latter was found to be more coherent and feasible.

- **Display Model:**
  Since the aim of retrieval is to find the target quickly, each display should capture as much information as possible about the target. This is formulated by maximizing the mutual information between target and answer (conditional on the query history). Motivated by this formulation, which is “optimal” in a sense to be described and yet computationally intractable, two heuristic solutions have been investigated. In addition, due to user fatigue and frustration, it is crucial to limit the expected number of iterations. In particular, it is necessary to find a suitable number of images to display at each iteration. If there are too many images, users are over-burdened in providing answers; if there are too few, the amount of information collected at each iteration is not sufficient to limit the search time.

- **Global Probability Model:**
  Once the answer and display models are fixed, the retrieval problem can be formulated in terms of Bayesian inference. The key random variables are the target and the answer sequence. Under the assumption of conditionally independent responses, the probability distribution on the target is updated after each response according to Bayes rule. This involves specifying a conditional answer probability distribution given each possible display.

The key question is whether there exists an image representation, specifically face image features and a corresponding metric, which at least roughly coheres with the subjective judgments made by users. And how does one measure the average degree of coherence between such a metric and the user’s response given different targets? These problems in coherence analysis connect signatures and metrics with the design of a relevance feedback system. Since subjectivity in real user responses is unavoidable [2, 15], the factor which can be controlled is how images are compared in the system. Hence, to make the whole system work, there is an inevitable adjustment of both metrics and signatures, and relevance feedback models, according to data collected about coherence.

3. SIGNATURES AND METRICS

In the system, images are represented by signatures and two images are compared with a metric on signature space.
What are the proper quantities for mental face retrieval? Different signatures have been used for different tasks such as face detection, face recognition and face retrieval by query-by-example. It is natural to try signatures developed for either face recognition, designed for identifying a particular person, or face retrieval with query-by-example, also aimed at searching a database for a particular target.

However, mental face retrieval is very different as it involves learning from the user, and the signature and metric should be chosen to somehow optimize collecting information. At minimum, there must be a degree of coherence between system matching and user matching. We summarize our analysis of coherence based on several subspace signatures.

3.1 Subspace methods

Subspace methods, such as principal component analysis (PCA) [18], Fisher’s discriminant analysis (FDA) [6, 1], the kernel version of PCA (KPCA) [20] and the kernel version of FDA (KFDA) [19, 12], are popular schemes for feature extraction in face recognition and face retrieval by example. The former two are linear methods and the latter two are nonlinear extensions of these.

Linear subspace methods map the image vector $X \in \mathbb{R}^M$ to another vector $Y \in \mathbb{R}^m$, $m << M$, of much lower dimension through a linear transform $Y = WX$ with a projection matrix $W$. Hence, the major issue is the choice of $W$. For PCA, the solution is a matrix composed of the eigenvectors of the covariance matrix based on centered samples, sorted by decreasing order of the eigenvalues. Hence, PCA maps the image vector to a subspace spanned by most representative bases. FDA aims at finding a $W$ that simultaneously maximizes the inter-class distance and minimizes the intra-class distance through maximization of Fisher’s criterion. With the corresponding projection operator $W$, each face image can be represented as a vector of weights, which are the inner products between the column vectors of the projection matrix and the image vector.

KPCA and KFDA are non-linear extensions of PCA and FDA, respectively, obtained through “kernel tricks”. Seen differently, they perform PCA and FDA in a much higher or even infinite-dimensional space through first mapping the image vector from $\mathbb{R}^M$ to $\mathbb{R}^F, F >> M$, by a nonlinear transformation. All the classes are expected to have a richer separation from each other. Since the various scatter matrices in PCA and FDA are involve only vector inner products, these items can be computed with Mercer kernels.

3.2 Metrics

The $L_1$ distance is adopted for subspace-based signatures. Since the signature is high dimensional and since the images are sparsely scattered in the high-dimensional space, the distribution of distances for a real image database is highly unstructured. To partly solve this problem, we performed normalization to all distances and the distance between images $i$ and $j$ is hereafter denoted as $d(i, j) \in [0, 1]$.

4. RELEVANCE FEEDBACK MACHINE

The search time (number of iterations until the target is located) is a random variable and our objective is to minimize some property of its distribution, such as the mean. (Here we are assuming there is a version of the mental image in the database and that the user will recognize it when displayed.) An intuitive and popular display strategy is to simply to show the images that are judged to be the most likely to be the target, as measured by the current posterior distribution on images in the database. However, such a strategy may not be efficient, i.e., minimize the search time, since the system might be stuck in local minimum due to poor estimation of the user’s intentions. That is, the posterior distribution may be very flat or peaked at the wrong images due to “noise” in the answers. Instead, we attempt to pick the most informative display based on mutual information. A Bayesian model is constructed and a conditional distribution is designed to interpret the user’s response given his target.

The following notation is convenient: $S$ indexes the images with positive integers $1, ..., N$; the target, a random variable, is denoted by $Y \in S$; $D_t \subset S$ is the set of $n$ images displayed at iteration $t$; $X_{D_t}$ denotes the response to the display $D_t$ with values $x_i \in A$; and $B_t = \{ x_{D_t} = x_i \}$ is the history of queries and answers through iteration $t$.

4.1 Display model

We wish to successively reduce the uncertainty about the target $Y$ by collecting information at each iteration of the display/answer process. Given a model for $X_{D_t}$ (see below), the efficiency of learning is determined by the amount of information collected. That is, the most informative query maximizes the decrease in uncertainty about $Y$. What is the same, the most informative query at iteration $t + 1$ is the display set $D_{t+1}$, for which the anticipated corresponding answer $X_{D_{t+1}}$ tells us the most about $Y$ given the current evidence $B_t$. Such quantities are well-studied in information theory.

Hence, we want to display the set $D \subset S$ that maximizes the mutual information between $Y$ and $X_D$ given the search history $B_t$. That is,

$$D_{t+1} = \arg \max_{D \subset S} I(Y; X_D | B_t)$$

where $H$ denotes entropy. Since $H(Y | B_t)$ is independent of $D$, Eqn.(1) is equivalent to

$$D_{t+1} = \arg \min_{D \subset S} H(Y | B_t, X_D)$$

(2)

That is to say, choosing $D \subset S$ to maximize the mutual information $I(Y; X_D | B_t)$ is equivalent to minimizing the conditional entropy $H(Y | B_t, X_D)$. It can be further shown (see Appendix A) that, under an ideal response model in which the user’s choice is perfectly consistent with the system metric, Eqn.(2) becomes

$$D_{t+1} = \arg \max_{D \subset S} H(X_D | B_t)$$

(3)

However, these combinatorial optimization problems are intractable. As a result, we resort to a heuristic solution motivated by attempting to maximize the objective function $H(X_D | B_t)$ corresponding to an ideal response [7]. Although not exact, the algorithm we employ has the advantage of speed and simplicity and seems to work well in practice.
4.2 Updating the posterior distribution

Suppose the prior probability of image \( k \) being the target is

\[
p_0(k) = P(Y = k), \quad k \in S
\]

After \( t \) queries, the posterior distribution of \( Y \) is the conditional law of \( Y \) given the query history \( B_t \):

\[
p_t(k) = P(Y = k | B_t), \quad k \in S
\]

Using Bayes rule, the posterior at iteration \( t + 1 \) is

\[
p_{t+1}(k) = P(Y = k | B_{t+1}) = \frac{P(Y = k) P(B_{t+1} | Y = k)}{\sum_{j=1}^{N} P(Y = j) P(B_{t+1} | Y = j)}
\]

\[
\propto p_t(k) P(B_{t+1} | Y = k)
\]

In our model, the basic assumption is that the answers \( X_{D_t} \) for different iterations are conditionally independent given \( Y \). Thus, we can compute the posterior at iteration \( t + 1 \) in terms of the posterior at iteration \( t \) as follows:

\[
p_{t+1}(k) \propto p_t(k) P(B_t, X_{D_{t+1}} = x_{t+1} | Y = k) = p_t(k) P(B_t | Y = k) P(X_{D_{t+1}} = x_{t+1} | Y = k)
\]

\[
\propto p_t(k) P(X_{D_{t+1}} = x_{t+1} | Y = k)
\]

where \( x_{t+1} \) is the answer for query \( X_{D_{t+1}} \). As a result, knowing the answer model is all we need to update the posterior.

4.3 Answer model

Given the target is image \( k \), the probability of answering \( x_t \) at iteration \( t \) is \( P(X_{D_t} = x_t | Y = k) \). Putting aside the functional form for the moment, we must carefully specify the set \( A \) of possible answers. Based on real experiments, we observe that humans can often identify cases in which two faces are very much alike and most people will hold similar opinions. However, there are many subjective factors. People have individual viewpoints due to differing life experience. A suitable mathematical formulation should reflect common behavior and try to avoid subtleties in behavior. Also, the answer set should be simple enough to promote some coherence between user responses and adopted metrics, and to avoid unduly burdening the user with complex decisions.

Based on these criteria, we adopted a comparative response, asking the user to make one global decision, namely choose the image which in his opinion is closest to the mental image. In this case, the set of possible answers is

\[
A = \{1, \ldots, n + 1, \ldots, 2n\}
\]

Suppose \( D_t = \{t_1, \ldots, t_n\} \) is the set of displayed images. Then \( x_{D_t} = i \) for \( i \in \{1, \ldots, n\} \) means that image \( t_i \) is considered the closest one to the target but is not actually the target, whereas \( x_{D_t} = i \) for \( i \in \{n + 1, \ldots, 2n\} \) means that image \( t_{i-n} \) is the target. Thus the conditional answer distribution given \( Y = k \) breaks into two cases depending on whether or not the display set contains \( k \).

In comparison with providing information about each displayed image, the comparative response provides less information at each iteration. But it is easier for the user of the system and easier to model in a way that coheres with human behavior. The answer set defined in Eqn.(7) proves superior to other definitions in practice, which is more suitable for mental face retrieval.

The form of the model we use is as follows:

If \( k \in D_t \), we have

\[
P(X_{D_t} = t_i | Y = k) = \begin{cases} 1 & \text{if } k = t_{i-n} \\ 0 & \text{otherwise} \end{cases}
\]

Otherwise, i.e., if \( k \not\in D_t \), then for \( i \in \{1, \ldots, n\} \):

\[
P(X_{D_t} = t_i | Y = k) = \frac{\phi(d(t_i, k))}{\sum_{j \in D_t} \phi(d(t_j, k))}.
\]

We make the natural assumption that the closer the distance in feature space between a displayed image \( t_i \) and the target \( Y \), the more likely is that image to be picked by user. Hence, we picked a monotonically decreasing function of distance as \( \phi() \), whose parameters are learned by Maximum Likelihood Estimation with collected real user responses (See Section 5.1).

5.0 ANALYSIS OF EXPERIMENTS

A web interface (see Fig.2) was designed to collect user response data for learning the function \( \phi \) and to evaluate the performance of the system. In each test, a target image is chosen according to the prior distribution \( \phi \) and displayed as in Fig.2 (a); groups of eight images are successively displayed in future pages and each time the user is required to pick the closest one to the given target (see Fig.2 (b)).

5.1 Face databases and user data

A series of subsets of the FERET face database is utilized in our experiments. The number of subjects (distinct individuals) is equal to the number of images, i.e., there is only one example of each subject in the database. In this way, the random variable \( Y \) is well-defined. FERET(A) contains all 1199 subjects in the FERET database. Since the majority of people in FERET(A) are Caucasian, the prior is ethnically nonuniform. As is well-known, ethnicity is a major semantic component in human decision-making based on comparisons. We tried to limit the ethnic influence by adopting other subsets. FERET(C) contains 808 Caucasian subjects and FERET(W) contains 327 Caucasian females. FERET(SB) contains 512 subjects with about equal numbers from three major ethnic categories (Asian, Black and Caucasian) and the two gender categories (male and female). The only difference between FERET(SB) and FERET(SB+F) is that the latter contains 19 face images which represent people familiar to our real users.

Experiments were designed to collect real user responses for both modeling and coherence analysis. The displayed images were picked either randomly or via the display model. Users were either researchers at the IMEDIA group at INRIA or engineers at the SAGEM Group. We collected over 4000 records of target/display/response.

5.2 Coherence analysis

In our work, we evaluate coherence in an average sense rather than for any specific user. Our analysis was restricted to the candidate pool of signatures mentioned earlier and to our answer scenario (comparative response).

Statistics of real user responses were collected for the various signatures and metrics introduced in Section 3. We
computed the cumulative distribution function for the rank of the user’s choice among the displayed images. This is possible in those simulations in which the mental image is chosen from among those in the database and is known to the system. In these cases, we can compute the distance from the target to each displayed image and therefore determine the rank of the image selected by the user: 1-st closest, 2-nd closest,...,8-th closest. Thus, the cumulative mass for the k’th closest represents the percentage of records in which the chosen image was among the k closest images to the target using the signature and metrics employed by the search engine. In particular, the ideal response model assumes all mass is concentrated on rank 1 and, accordingly, the cumulative probability is 1 for all 8 ranks. In view of Table 1, this is obviously far from reality. In general, the larger the value of the cumulative probability, the higher is the coherence with the signatures and metric. Examples of results on FERET(SB) are shown in Table 1.

In Table 1, we can observe that the systems based on the various signatures on FERET(SB) are all roughly equally coherent with real users. We observe similar results in our experiments on other databases. We can also observe that no representation is highly adapted to human decision-making. In the experiments for testing relevance feedback model, we fix the $L_1$ metric on KFDA features for simplicity.

### 5.3 Performance analysis

The performance of the system is evaluated using statistics based on the iteration number $T \in \{1, ..., N\}$, the number of image groups displayed before the target is identified. In each experiment with an individual user or group of users, $N_T$ complete searches were performed, resulting in $N_T$ samples of $T$ denoted $k_1, k_2, ..., k_{N_T}$. Performance was based on the average $E(T)$ and cumulative probability $F_T(t)$, estimated respectively as

$$E(T) = \frac{1}{N_T} \sum_{i=1}^{N_T} k_i$$

and

$$F_T(t) = \frac{\# \{i | T = k_i \leq t, i = 1, 2, ..., N_T\}}{N_T}$$

Evidently, a good algorithm has a small $E(T)$ and the graph of cumulative distribution $F_T(t)$ grows rapidly as iteration number increases.

There are many factors whose effect on performance require exploration, including the influence of coherence, the number of displayed images and the size of the database. Some results have been reported in our previous paper [7].

Table 1: Coherence Analysis

<table>
<thead>
<tr>
<th>RANK</th>
<th>PCA</th>
<th>LDA</th>
<th>KPCA</th>
<th>KLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.21</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>0.37</td>
<td>0.37</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>3</td>
<td>0.50</td>
<td>0.49</td>
<td>0.49</td>
<td>0.48</td>
</tr>
<tr>
<td>4</td>
<td>0.61</td>
<td>0.61</td>
<td>0.60</td>
<td>0.58</td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.73</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>0.81</td>
<td>0.82</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>7</td>
<td>0.92</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Figure 2: Web interface for comparative answer scheme: (a) target page; (b) display page.
that the average iteration number grows slowly with the size of the database, which is encouraging for applications, such as criminal identification, in which very large databases are involved. More extensive results than those reported in [7] are now reported.

- **Influence of the number of displayed images:**
  We vary $n$, the number of images shown at each iteration, and fix all other conditions. The results with simulations ($N_T = 100$) are shown in Fig.3, in which the database is FERET(A). By simulation we mean that the answers are actually sampled from the answer model, i.e., a target is selected at random and, at each iteration, the display model algorithm generates $n$ images but the answer is a random sample from the conditional distribution given the target. This is obviously an idealized scenario, but allows us to perform many experiments and obtain qualitative results. The graphs of $F_T(t)$ are shown for $n = 4, 8, 12, 16$. In addition to $E(T)$, we show $n \times E(T)$ (average total number of images shown before finding the target).

As expected, it can be observed that $E(T)$ decreases as $n$ increases. However, we also observe that the total number of images displayed increases with $n$. In real experiments, as $n$ grows, at each iteration users will be required to analyze more images, i.e., do perform more mental matching. On the other hand, if $n$ is small, users must view more total pages before locating their target. Hence, there is a trade-off between the number of images shown in each display and number of iterations required. As a compromise, we fix $n = 8$ in all the ensuing experiments.

- **Sensitivity to the particular target:**

  The images are represented in a high-dimensional signature space, and the distribution of distances from image to image is rather irregular. Thus, some targets are located in dense areas in that there are many other images at small distances, whereas other targets are relatively isolated. It is not realistic to evaluate this effect analytically. Instead, we performed a rough assessment using the following measure of isolation:

  $$d^*(k) = \frac{\sum_{i=1}^{N} d(k, i) - d_{\text{min}}}{d_{\text{max}} - d_{\text{min}}}$$

  where $d_{\text{max}} = \max_{j=1,...,N} \sum_{i=1}^{N} d(j, i)$ and $d_{\text{min}} = \min_{j=1,...,N} \sum_{i=1}^{N} d(j, i)$. (The metric $d$ is not normalized.) Thus the larger $d^*(k)$, the farther away on average is image $k$ from the other images in the database.

  We measured the effect of $d^*$ on performance. The tests is done on FERET(A). Three face images with $d^* = 0.0, 0.5$ and 1.0 were chosen as the target in simulations, repeated 100 times in each case. The results are shown in Fig.4. We can observe that when $d^*$ is large, that is, the target is geometrically isolated in the database, the search is much more efficient, whereas when $d^* = 0$, so that there are many other images very near the target, the average iteration number $E(T)$ is larger. Notice in Fig.4 that the average iteration numbers differ by 10 units for the two extreme cases. However, it is not clear this large effect would carry over to experiments with real users. This remains to be investigated.

- **Tests with real users:**

  In simulations, the model performs well under various conditions. This is not surprising in view of the fact that the responses are actually sampled from the model; hence some measure of coherence is hard-wired. However, what is the performance with real users?
In the first experiment with humans and the FERET database, the target image is continuously displayed. The reason is that our users are not “familiar” with a target randomly chosen from this database and can not be expected to have a reasonable “mental image.” The experiment involved nine researchers in the IMEDIA project and the database FERET(SB). We collected a total of \(N_T = 48\) complete searches.

Of course the scenario above is not real “mental face retrieval.” On the other hand, it is also not realistic to ask a user to “remember” the face of somebody with whom the user has no prior interaction; the answers are likely to be too random. To address this problem, we add face images of people familiar to the users to the database and use these additional images as targets. The target is then not persistently shown as above and the responses of the users are entirely based on memory. This experiment was performed by 22 INRIA researchers using the FERET(SB+F) database. A total of \(N_T = 78\) full searches were performed.

The graphs of \(F_T(t)\) for both experiments are shown in Fig.5. The performance under the two settings is actually very similar. For purposes of comparison, note that for a totally random display, the cumulative distribution of \(T\) is \(F_T(t) = \frac{t}{40}\). In Table 2, we show \(F_T(t)\) for selected values of \(t\) for each of the experiments and for random display. It can be concluded that the proposed model far out-performs a random display with real users, with or without showing the target. These results establish the feasibility of the mental face retrieval system for moderately-sized databases.

There is one more point about the statistical analysis of the experiments with real users that should be mentioned. As seen in Fig.5, the value of the cumulative distribution function does not reach one at iteration 40; in other words, some search times exceeded 40. Of course the mean values are based on all the data. In practice, extensive searches are exhausting to people, decisions become increasingly arbitrary and it is not realistic to continue iterating. If we define those tests with iteration number above some value (say 40 for example) as “failed”, then the average iteration number of “successful” tests will be smaller than those shown in the Figures. This may be a more suitable way to measure performance: provide the probability of a failed search together with statistics on the successful ones.

### 6. CONCLUSIONS

An interactive system of mental face retrieval is proposed using the framework of relevance feedback and Bayesian modeling. The scenario is based on purely comparative, subjective judgments about distances between a displayed image and an image “in mind”. Assuming there is a unique representation of the mental image in the database, the responses of real users are modeled as a conditional probability distribution given the target. The choice of which images to display is motivated by maximizing mutual information assuming an ideal user whose selections are in perfect agreement with the underlying system metric.

Experiments demonstrate that mental face retrieval by information-driven feedback works well in simulations and yields promising results with real users. In particular, the experiments show stability of performance with respect to the size of database and the choice of signatures. Real users are able to locate their target in fewer than twenty iterations 74.4% of the time with databases of order 1000.

In order to significantly improve the system, it is likely that a more profound “coherence analysis” is necessary, allowing this criterion to drive the selection of image features and metrics. In addition, since the coherence of particular signatures and metrics will likely depend on the content and size of the database, more work with larger databases is necessary to obtain a practical system for certain applications. Finally, introducing semantic annotation may be necessary to obtain realistic search times with very large databases.

### 7. ACKNOWLEDGMENTS

We acknowledge many useful suggestions in designing the model and developing the interface from Mr J.-P. Chieze, Dr. F. Fleuret, Dr. S. Hichem of IMDEIA, INRIA and Mr. P. Welti of the SAGEM group.

### 8. REFERENCES


**APPENDIX**

### A. THE DISPLAY OBJECTIVE FUNCTION UNDER IDEAL RESPONSE

In the case of an ideal response, for any \( D = \{ s_1, s_2, \ldots, s_n \} \subset S \), the answer model is of the form

\[
P(X_D = i| Y = k) = \begin{cases} 1 & \text{if } i \in \{1, \ldots, n\} \text{ and } s_i \neq k \\
               d(k, s_i) & \text{for all } j = 1, \ldots, n \\
               0 & \text{otherwise}.
\end{cases}
\]

Then

\[
H(X_D|Y) = \sum_{i=1}^{N} P(Y = i) H(X_D|Y = i) = 0
\]

Moreover, since \( Y \) determines \( X_D \),

\[
H(X_D|B_i, Y) = 0
\]

That is to say

\[
H(Y, X_D|B_i) = H(Y|B_i)
\]

By applying the above equation,

\[
H(Y|B_i, X_D) = H(Y, X_D|B_i) - H(X_D|B_i)
\]

and hence, choosing \( D \subset S \) to minimize the conditional entropy \( H(Y|B_i, X_D) \) is equivalent to maximizing the entropy \( H(X_D|B_i) \) in the case of an ideal response. Thus, the optimal display for an ideal response is

\[
D_{i+1} = \arg \max_{D \subset S} H(X_D|B_i)
\]