Experiments in Mental Face Retrieval

Yuchun Fang[1] and Donald Geman[1,2]

[1] IMEDIA Project, INRIA Rocquencourt yuchun.fang@inria.fr
[2] Dept. Applied Math. and Stat., Johns Hopkins University geman@jhu.edu

Abstract. We propose a relevance feedback system for retrieving a mental face picture from a large image database. This scenario differs from standard image retrieval since the target image exists only in the mind of the user, who responds to a sequence of machine-generated queries designed to display the person in mind as quickly as possible. At each iteration the user declares which of several displayed faces is “closest” to his target. The central limiting factor is the “semantic gap” between the standard intensity-based features which index the images in the database and the higher-level representation in the mind of the user which drives his answers. We explore a Bayesian, information-theoretic framework for choosing which images to display and for modeling the response of the user. The challenge is to account for psycho-visual factors and sources of variability in human decision-making. We present experiments with real users which illustrate and validate the proposed algorithms.

Keywords: relevance feedback, mental image retrieval, Bayesian inference

1 Introduction

Traditional image retrieval is based on “query-by-example”: starting from an actual image, the objective is to find the images in the database which are visually similar to the query image. Striking results are obtained in special domains, e.g., in comparing paintings, plants and landscapes using the IKONA system [1].

However, in many cases of interest there is no physical example to serve as the query image [2]. Instead, knowledge about the target is based entirely on the subjective impressions and opinions of the user. In other words, the standard query image is replaced by a “mental image”. To be concrete, we shall concentrate throughout on face images, although all the algorithms we develop could be applied in other domains, for instance to images of clothes, houses, funitures or paintings. Mental face retrieval has extensive applications in security, e-business, web-based browsing and other areas. Here, as the realization of a study conducted jointly with the SAGEM group, we propose a system for retrieving a mental face image using Bayesian inference and relevance feedback. It is based on an interactive process designed to incrementally obtain knowledge about the target from the responses of the user to a series of multiple choice questions. The objective is to minimize the number of iterations until a face is displayed whose identity corresponds to the mental image.

Thus relevance feedback refers to a series of queries and answers. The query is simply a set of displayed images from the database. The answer is the feedback provided by the user. Usually, the opinions or impressions of the user concerning both his target and the displayed images are of high-level, semantic nature, and hence "mental matching" involves human memory, perception and opinions. On the other hand, the representation of the images in the database is generally based on low-level signatures rather than semantic content. This "semantic gap" greatly complicates the task. Indeed the face recognition problem, which is arguably easier, remains largely unsolved, at least with large databases.

Still, if the display and answer models are constructed to explicitly address the issue of coherence, it is possible to incrementally obtain knowledge about the target image. The accumulation of information is represented by an evolving probability distribution over the database, whose entropy is hopefully diminishing (although not monotonically) as information is acquired from the answers. This process of alternating between query and answer is iterated until the user recognizes one of the displayed images as his target, at which point the search terminates. The two primary challenges in mental picture retrieval are then deciding which images to display at each iteration (the "display model") and accounting for the difference between mental matching and signature-based matching (i.e., between mental and feature-based metrics) in designing the conditional probability distribution for the answers given the target (the "answer model").

In our framework, both the target and answers to queries are treated as random variables; the probability distribution of the target evolves over time based on the accumulated evidence from the user's responses. A natural choice for the images to display at each iteration is then the set which maximizes the mutual information between the target and response given all previous answers. As this optimization problem is intractable, a heuristic solution is proposed based on an "ideal" answer model which puts the user and system in synchrony. In addition, in order to find image representations which cohere as much as possible with human decision making, we compare several traditional face recognition signatures. Based on this analysis as well as data collected from human responses, in particular declaring which among a set of displayed images is "closest" to a given target, an answer model is designed for a comparative response. The feasibility of the whole system is demonstrated by estimating mean search times and other summary statistics from mental retrieval experiments with real users.

Whereas there has been considerable work done on face retrieval in the standard setting of query-by-example [6, 4], little has been reported in the case of mental images. Navarret et al. proposed an algorithm based on self- organizing maps [8]; see also the work on "retrieval of ambiguous target" in [9]. Of course, there are many articles on relevance feedback [13], however, most of them involve "category search", which is different from "target search" in the case of mental face retrieval. In our view, the benchmark work on "target search" for mental images is Cox et al [3]; see also the model proposed by Geman and Moquet [5] for the toy application of mental polygon retrieval. By concentrating on the interactive process and specializing to target search and pairwise comparison tests, the authors in these studies were able to develop ties with Bayesian inference and information theory. However, the answer model in [3], basically a blurring of the actual metric used by the system in comparing two images, is not sufficiently powerful to deal with face retrieval. Moreover, pairwise comparison search is not practical with large image databases. We believe our work constitutes the first comprehensive study of mental face retrieval, both in terms of mathematical foundations and experiments with real users.

The remainder of the paper is organized as follows. The formulation of the problem in terms of Bayesian relevance feedback is described in Section 2. The answer model and display model are explained in detail in Sections 3 and 4 respectively. In Section 5, we discuss signature extraction and analyze the coherence issue. Experimental results are presented in Section 6.

2 Bayesian Relevance Feedback Model

In the framework we propose, mental image retrieval will depend on solving two difficult tasks:

– **A Modeling Problem:** Discovering answer models which match human behavior;
– **An Optimization Problem:** Discovering approximations to the optimal query.

Suppose there are N images in the database $\mathcal{S}$, say $I_1, ..., I_N$. For simplicity, we will identify $\mathcal{S}$ with the index set $\{1, 2, ..., N\}$. One image in the database, Y, is the "target", i.e., the variation on the mental picture assumed to belong $\mathcal{S}$. In the stochastic formulation, Y is a random variable with some initial distribution

$$p_0(k) = P(Y = k), \ \ k \in \mathcal{S}.$$

Information about Y is collected from a series of queries. Each query involves two quantities: a subset $\mathcal{D} \subset \mathcal{S}$ of n displayed images and the response of the user, denoted by $X_{\mathcal{D}}$ and taking values in a set $\mathcal{A}$. Obviously $n \ll N$; the choices for n and $\mathcal{A}$ are important issues which will be discussed in the following sections.

The feedback from the user up to time (or iteration) $t = 1, 2, ...,$, is then

$$B_t = \bigcap_{i=1}^{t} \{X_{\mathcal{D}_i} = x_i\}$$

where $\mathcal{D}_i$ is the display at time i and x_i is the user's response. This is the history of queries and answers during the first t iterations.

We wish to compute and update the posterior distribution,

$$p_t(k) = P(Y = k|B_t), \ \ k \in \mathcal{S},$$

the probability that image k is the target after t iterations. First, however, we must specify the joint distribution of Y and the observations $\{X_{\mathcal{D}_1}, ..., X_{\mathcal{D}_t}\}$.

The posterior p_t is then computed in the usual way. As in previous work, we are going to assume the answers to the queries are conditionally independent given the target Y. This is not an unreasonable assumption in practice. It follows that

$$P(B_t|Y=k) = \prod_{i=1}^{t} P(X_{D_i} = x_i|Y=k).$$

The conditional response distribution, $P(X_{\mathcal{D}} = x|Y=k)$ is what we call the "answer model."

Updating the posterior is now easy:

$$\begin{aligned} p_{t+1}(k) &= P(Y=k|B_{t+1}) \\ &= P(Y=k|B_t, X_{D_{t+1}} = x_{t+1}) \\ &\propto p_t(k) P(X_{\mathcal{D}_{t+1}} = x_{t+1}|Y=k) \end{aligned}$$

In other words, updating $p_t(k)$ merely involves multiplying by the new evidence $P(X_{\mathcal{D}_{t+1}} = x_{t+1}|Y=k)$ and re-normalizing.

3 Answer Model

Designing $P(X_{\mathcal{D}} = x|Y=k)$ involves two primary decisions: determining the set of possible responses $x \in \mathcal{A}$ and capturing the behavior of a user who has image k in mind and is presented with the images in D and asked to respond. This specification inevitably relies on the metric in the signature space, denoted by d. More details about this metric is introduced in Section 5.

There are many possible choices for $\mathcal{A}$. In all cases, the target is identified if present, so let us assume that $Y \notin \mathcal{D}$. One could ask the user to supply a rather precise measure of the degree of similarity between each displayed image and his target. Somewhat less demanding, one could solicit a rough label for each displayed image, such as "relevant" or "not relevant". We have adopted a still simpler scheme in which the user is simply asked to select the image which is "closest" to his target. The price for simplicity is of course a decrease in the amount of information conveyed, and hence in the reduction of uncertainty about Y. Nonetheless, in our experiments, this model proved to be the most practical, both mathematically and in terms of user psychology. It does not unduly burden the user with complex decision-making, nor require any specific knowledge of image representation, and it provides a natural way of bringing the stored metrics into play. To make this precise, assume $\mathcal{D} = \{s_1, ..., s_n\}$ and set

$$\mathcal{A} = \{1, ...n, n+1, ..., 2n\} \tag{1}$$

For $i \in \{1, ..., n\}$, the response $X_{\mathcal{D}} = i$ signifies that image s_i is not the target but, *in the opinion of the user*, is the one closest to his target. Response $i \in \{n+1, ..., 2n\}$ signifies that image s_{i-n} is the target.

By definition of such comparative answer, if $k \in \mathcal{D}$, we have

$$P(X_{\mathcal{D}} = i | Y = k) = \begin{cases} 1 \text{ if } k = s_{i-n} \\ 0 \text{ otherwise} \end{cases}$$

Otherwise, i.e., if $k \notin \mathcal{D}$, then for $i \in \{1, ..., n\}$:

$$P(X_{\mathcal{D}} = i | Y = k) = \frac{\phi(d(s_i, k))}{\sum\limits_{s_j \in \mathcal{D}} \phi(d(s_j, k))} \tag{2}$$

Ideally, the closer the image $s_i \in \mathcal{D}$ is to k in the stored metric, the more likely the user is to choose it. Hence, we take ϕ to be monotonically decreasing. In our experiments, we adopt a parametric form for ϕ and learn the parameters from real data (collected user responses) by maximum likelihood estimation.

4 Display Model

One straightforward solution to determine $\mathcal{D}_t$, the n images displayed at iteration t, is to pick the n images which are most likely under the posterior distribution p_t. However, this simple strategy is far from optimal in terms of minimizing the average search time (our ultimate goal) except near the end of efficient searches, when p_t is highly concentrated. Instead, as in other work, we adopt the powerful (and time-independent) strategy of choosing $\mathcal{D}_{t+1}$ to minimize the uncertainty of the target given the search history B_t and new answer $X_{\mathcal{D}_{t+1}}$, measuring uncertainty by entropy:

$$\mathcal{D}_{t+1} = \arg\min_{\mathcal{D} \subset \mathcal{S}} H(Y | B_t, X_{\mathcal{D}}) \tag{3}$$

Since the entropy $H(Y|B_t)$ is independent of $\mathcal{D}$, Eqn.(3) is equivalent to maximizing the conditional mutual information between Y and $X_{\mathcal{D}}$ given B_t:

$$\mathcal{D}_{t+1} = \arg\max_{\mathcal{D} \subset \mathcal{S}} I(Y; X_{\mathcal{D}} | B_t) \tag{4}$$

The mutual information is then computed relative to the joint distribution determined by the answer model and the current posterior on the target.

4.1 Heuristic Solution

The minimization in Eqn.(3) is, unfortunately, a virtually intractable combinatorial optimization problem since there are $\binom{N}{n}$ choices for $\mathcal{D} \subset \mathcal{S}$. (Discarding images already displayed makes little difference.) The algorithm we use is based on an approximation to the corresponding optimization problem resulting from the choice of an ideal answer model under which the user selects the displayed image actually closest to his target using the system metric (or of course selects the target itself if present). Since Y determines X_{D+1}, it is easy to see that Eqn. (3) is equivalent

$$\mathcal{D}_{t+1} = \arg\max_{\mathcal{D} \subset \mathcal{S}} H(X_{\mathcal{D}} | B_t) \tag{5}$$

However, there is a natural heuristic for this combinatorial optimization problem.

Roughly speaking, since entropy is maximized at the uniform distribution, and ignoring the case in which the target belongs to $\mathcal{D}$, we want to choose n images, call them $\{s_1, ..., s_n\}$, such that the Voronoi partition has cells of almost equal mass under the posterior. A sequential, heuristic solution is then given by the following algorithm:

1. The candidate set $\mathcal{C}_1$ for s_1 consists of all images not previously displayed through iteration t.
2. Select s_1 to be the image $k \in \mathcal{C}_1$ which maximizes $p_t(k)$.
3. Order the images in $\mathcal{C}_1$ according to their distance to s_1. Add these one-by-one to a cluster initialized by $\{s_1\}$ until the mass of the cluster under p_t reaches $\frac{1}{n}$.
4. Define the candidate set $\mathcal{C}_2$ for choosing s_2, by removing the cluster from $\mathcal{C}_1$.
5. Select s_2 to be the image $k \in \mathcal{C}_2$ which maximizes $p_t(k)$.
6. Divide all the images in $\mathcal{C}_1$ into two groups: those closest to s_1 and those closest to s_2. Order the distances in the first group (respectively, second group) according to their distance to s_1 (resp. s_2) and repeat the agglomeration procedure in step 3 relative to both s_1 and s_2. This results two clusters "centered around" s_1 and s_2, each with mass approximately $\frac{1}{n}$.
7. Continue in this way until $\{s_1, ..., s_n\}$ are chosen.

Although there is no guarantee to maximize entropy in Eqn (5), this heuristic solution is fast, simple and achieves good performance in practice.

5 Signatures, Metrics and Coherence

Given our emphasis on retrieving mental images of faces, it would seem natural to use signatures developed for face recognition and face retrieval with query-by-example. As a result, we have analyzed several subspace-based signatures applied in these areas, such as principle component analysis (PCA) [11], Fisher's discriminant analysis (FDA) [10], and the kernel versions of PCA (KPCA) [12] and FDA (KFDA) [7]. It should be emphasized however, that in face recognition and retrieval, the target image is available and hence its signature can be computed and directly compared with the signatures of other stored images. In particular, there is no guarantee that effective signatures for face recognition will also prove useful in mental retrieval.

We adopt the L_1 distance (Performance with L_2 is roughly the same) with normalization by size of database and order of value in database as our metric. One reason for the normalization is that standard signatures of the images in $\mathcal{S}$ are sparsely scattered in a high-dimensional Euclidean space and there is enormous variability among the distances between image pairs. Normalizing the distance using the order statistics ameliorates this problem.

We investigated the coherence between mental matching and metric-based matching by collecting responses from various individuals. All the experiments in this paper utilize subsets of the FERET database. Since the majority of people in the FERET database are Caucasian, and since the response of most people is heavily biased by semantics, we used the FERET(SB) (see Table 1) in the coherence experiment. In FERET(SB), the distribution of ethnic (Asian, Black and Caucasian) and gender categories (female and male) is roughly uniform. Each data item consists of a triple $(Y, \mathcal{D}, X_{\mathcal{D}})$ corresponding to a target, set of displayed images and user response. The targets were sampled at random from S and the number of displayed images is set to $n = 8$. (Using many fewer or many more has adverse consequences with real users.) The answers are comparative, as described in Section 3. Nine individuals (in the INRIA labs) produced a total of 989 data items (records). Statistics were collected on the rank of the user's choice in terms of the L_1 distance between each display and the target. An example experiment is shown in Fig.1, which compares PCA and KFDA under the L_1 metric; both the density of rank and its cumulative distribution are shown. These two signatures perform about the same. Neither can be said to be highly coherent with mental matching as the probability that the user selects the closest image is only roughly 0.2. Nonetheless, reasonable search times are obtained; see Section 6. Similar results are observed in other signature spaces. Henceforth, we fix our distance to be the L_1 metric on the KFDA image representation.

Table 1. Face databases used in experiments

NAME	#Subjects	#Images	Composition
FERET(A)	1199	1199	All FERET images
FERET(C)	808	808	Caucasian subset
FERET(SB)	512	512	Semantically balanced subset
FERET(W)	327	327	Caucasian female subset
FERET(SB+F)	531	531	FERET(SB)+ 19 extra (familiar) faces

6 Experiments in Relevance Feedback

The web interface in the experiments is shown in Fig.2. Let T denote the number of iterations (query/response) until the target appears among the displayed images. Given M tests (full searches), we estimate $E(T)$, the mean of T, and $P(T \leq t)$, the (cumulative) distribution of T by their empirical statistics. That is, if the M tests results in search times $T_1, ..., T_M$, then $E(T) = \frac{1}{M} \sum_{m=1}^{M} T_m$ and $P(T \leq t) = \frac{\#\{1 \leq m \leq M | T_m \leq t\}}{M}$. Evidently, we seek small values of $E(T)$ and cumulative distributions $P(T \leq t)$ which climb as fast as possible.

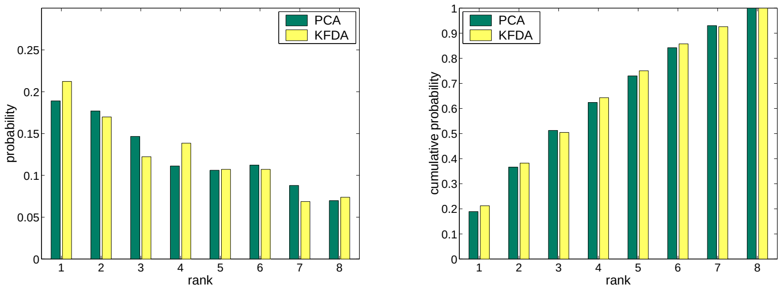


Fig. 1. Results comparing PCA and KFDA. Left: The estimated probability that the user selects the m'th closest image to his target according to the distance in signature space; Right: The cumulative distribution function.

Experiment I: Influence of the Answer Model

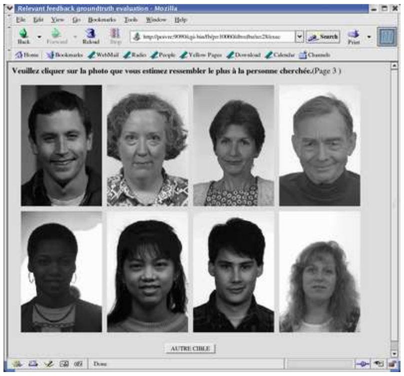
We designed answer models with varying degrees of "noise" in the sense of how well decisions cohere with the actual metric on signatures. For answer model defined in Eqn.(2), synchronization is controlled by the function $\phi(d)$ where $d = d(s_i, Y)$, the distance between the "i'th" displayed image and the actual target Y. The more rapidly $\phi(d)$ decreases (as d increases) the more likely is the user's answer to cohere with the signature metric. We did simulations with four answer models, meaning the answers are generated by sampling from the model. The response of the "ideal user" is always perfectly coherent with metric, i.e., $P(X_{\mathcal{D}} = i|Y = k) = 1$ if $d(s_i, k) < d(s_j, k)$ for all $s_i, s_j \in \mathcal{D}, i \neq j$. This represents the optimal performance obtainable. The other extreme is a random response ($\phi(d) \equiv const$); every displayed image is equally likely to be chosen regardless of its distance to the target. Two cases in between, and far more realistic, are $\phi(d) = \frac{1}{d}$ and $\phi(d) = 1 - d$; the former is evidently more coherent than the latter. One simulation on FERET(A) (see Table 1) with $M = 100$ is shown in Fig.3. In addition to the four (estimated) distribution function, the (estimated) mean search time is listed in the legend box. Clearly the degree of coherence with the metric on signatures characterizes the performance.

Experiment II: Sensitivity to the Size of the Database

To measure the effect of $N = |\mathcal{S}|$, we used databases of increasing size: FERET(W) ($N = 327$), FERET(SB) ($N = 512$), FERET(C) ($N = 808$) and FERET(A) ($N = 1199$)(see Table 1). The curve in Fig.4 shows the variation of $E(T)$ with N. The average search time grows slowly with N, roughly logarithmically.

Experiment III: Performance with Real Users

Tests with real users and a standard research database such as FERET is problematic since the user is not familiar with the people represented in the database. Of course one can select an image at random and ask the user to "memorize it" for few seconds, but this does not provide for a realistic scenario. Instead, we add images of the faces of familiar people to the database and select these as the targets for our experiments with mental image retrieval. The results shown in Fig.5 are based on $M = 78$ complete searches conducted by 22 INRIA researchers using the FERET(SB+F) database (see Table 1). For comparison, we show a simulation under the same experimental setting (i.e., same answer and display models) as well as the distribution corresponding to random display. In this case, it is easy to see that the cumulative distribution is linear, $P(T \leq t) = \frac{t \times n}{N}$, and the $E(T) = \frac{T_{max}(1+T_{max})n}{2N}$, where T_{max} is the maximum number of iterations possible. The answer model uses a ϕ-function with the free parameters estimated by maximum likelihood. Obviously, the proposed model far out-performs a random response. More importantly, the absolute performance is quite reasonable, with a mean search time of $E(T) \approx 14.7$ iterations and target recovery in fewer than ten iterations in approximately one-half the searches. Fine-tuning the model, such as finding metrics and signatures more coherent with mental matching, would likely result in further improvements.


Relevant feedback groundtruth evaluation - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop Search Print
Home Bookmarks WebMail Radio People Yellow Pages Download Calendar Channels
AUTRE CIBLE
Done

Veuillez cliquer sur la photo que vous estimez ressembler le plus à la personne cherchée.(Page 3)

Fig. 2. The interface for experiments

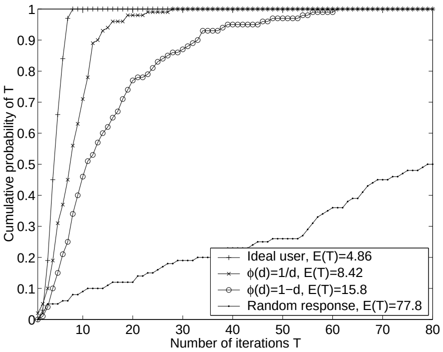


Fig. 3. The influence of the answer model

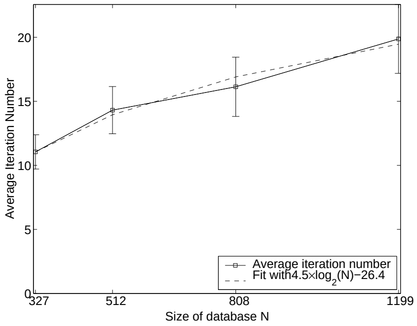


Fig. 4. The influence of N. The box plots are 95% confidence intervals.

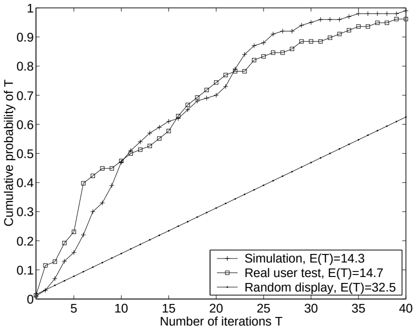


Fig. 5. Experiments in mental image retrieval with familiar faces as targets

7 Conclusions and Future Work

We have constructed a Bayesian model for mental face retrieval within the framework of relevance feedback. In deciding which faces to display to the user to match to the mental picture, a heuristic solution has been proposed based on the maximization of mutual information. The design of the answer model is motivated by the need to account for the variability in the responses of actual users and the lack of a strong correlation between the basis for mental matching and how images are compared using standard metrics on standard image features. The performance of the system is validated in both simulations and in experiments with real user tests, which demonstrate the feasibility of the proposed model. Improvements are likely to result from metrics and features more adapted to human decision making. Some degree of semantic annotation would also increase efficiency, especially with much larger databases.

Acknowledgment: We are grateful to Ms. N. Boujemaa for helping us to develop the feedback model. We also acknowledge many useful suggestions from Mr. P. Welti of the SAGEM group and the aide of Mr. J-P Chieze in the design of the user inteface.

References

1. Boujemaa, N., Fauqeur, J., Ferecatu, M., and et al. Ikona: Interactive generic and specific image retrieval. In *Intern. Workshop MMCBIR'2001* (2001).
2. Boujemaa, N., Fauqeur, J., and Gouet, V. *What's beyond query by example?* Springer Verlag, 2004.
3. Cox, I., Miller, M., Minka, T., Papathomas, T., and Yianilos, P. The bayesian image retrieval system, pichunter: theory, implementation and psychological experiments. *IEEE Trans. Image Processing 9* (2000), 20–37.
4. Eickeler, S. Face database retrieval using pseudo 2d hidden markov models. In *Proc. IEEE FG2002* (2002).
5. Geman, D., and Moquet, R. Q & a models for interactive search. Tech. rep., Dept. of Mathematics and Statistics, University of Massachusetts, 2001.
6. Liu, C., and Wechsler, H. Robust coding schemes for indexing and retrieval from large face databases. *IEEE Trans. Image Processing 9* (2000), 132–137.
7. Liu, Q., Huang, R., Lu, H., and Ma, S. Face recognition using kernel based fisher discriminant analysis. In *Proc. IEEE FG2002* (2002), pp. 197–201.
8. Navarrete, P., and Ruiz-del Solar, J. Interactive face retrieval using self-organizing maps. In *Proc. Int. Joint Conf. on Neural Networks* (2002).
9. Oda, M. Interactive search method for ambiguous target image. In *Proc. IEEE Intern. Workshop IDB-MMS'96* (1996), pp. 194–201.
10. Swets, D. L., and Weng, J. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. PAMI 18*, 8 (1996), 831–836.
11. Turk, M., and Pentland, A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience 3*, 1 (1991), 71–86.
12. Yang, M., Ahuja, N., and Kriegman, D. Face recognition using kernel eigenfaces. In *Proc. IEEE ICIP* (2000), vol. 1, pp. 37–40.
13. Zhou, X. S., and Huang, T. S. Relevance feedback for image retrieval: a comprehensive review. *ACM Multimedia Systems Journal 8*, 6 (2003), 536–544.