The Mug-Shot Search Problem

Ellie Baker and Margo Seltzer
Division of Engineering and Applied Sciences
Harvard University
40 Oxford Street
Cambridge, MA 02138
e-mail: ellie@eecs.harvard.edu

Abstract

Mug-shot search is the classic example of the general problem of searching a large facial image database when starting out with only a mental image of the sought-after face. We have implemented a prototype content-based image retrieval system that integrates composite face creation methods with a face-recognition technique (Eigenfaces) so that a user can both create faces and search for them automatically in a database.

Although the Eigenface method has been studied extensively for its ability to perform face identification tasks (in which the input to the system is an on-line facial image to identify), little research has been done to determine how effective it is when applied to the mug shot search problem (in which there is no on-line input image at the outset, and in which the task is similarity retrieval rather than face-recognition). With our prototype system, we have conducted a pilot user study that examines the usefulness of Eigenfaces applied to this problem. The study shows that the Eigenface method, though helpful, is an imperfect model of human perception of similarity between faces. Using a novel evaluation methodology, we have made progress at identifying specific search strategies that, given an imperfect correlation between the system and human similarity metrics, use whatever correlation does exist to the best advantage. The study also indicates that the use of facial composites as query images is advantageous compared to restricting users to database images for their queries.

Keywords: “Content-Based Image Retrieval,” “Face Recognition,” “Eigenfaces,” “Identikit”

1 Introduction

A tremendous amount of on-line image data is currently becoming available, but finding a particular image in a very large database of images is still a difficult problem. Images can be annotated with descriptive text and retrieved with traditional text-based query methods, but creating annotations requires substantial manual effort and the annotations are rarely sufficient to capture fully the content of an image. *Content-based image retrieval* systems [GR95] attempt to overcome the problems of text-based searching by permitting a user to specify image attributes in ways that are more direct and natural than the English language-like specifications required by traditional databases. One powerful approach is to let the user express the query with images rather than words (i.e., an *image-based query*). The system automatically compares the query image to those in the database and the most similar ones are retrieved. This approach, which can be studied independently from, and used in conjunction with, text-based methods, is our general focus.

Our research addresses the specific problem of content-based retrieval in large facial image databases. In particular, we assume a user begins the query process with only a mental image of a sought-after face. We refer to this as “the mug-shot search problem.” Since the database is large, manually inspecting every image is impractical. In fact, though the search space is finite (so theoretically one might be able to spend the time required to look through all of it), a sequential search can still fail because the user’s mental image can degrade or become confused as a result of viewing a large number of faces [CJ91]. Hence, we seek a query method that minimizes the number of image inspections required to find the face (or to determine that it is unlikely to be present in the database).

2 Background

The Photobook content-based image retrieval system [PPS94] provides one solution to this problem. Photobook uses Eigenfaces [TP91, KS90], a face identification technique based on principal component analysis (PCA). Using PCA, images consisting of N by N pixel intensity values are compressed from the high dimensional space of the N^2 pixel values to the much lower dimensional space of a small set of basis vectors called *eigenfaces*. Each face in the database can be roughly reconstructed as a weighted sum of the eigenfaces. The

weights are used to determine the distance (e.g., euclidean) between images. Distance from a query image is used to specify a sort order on the database. Typically, the user selects an initial query image from a small set of images selected randomly from the database. The system sorts the database relative to the query and presents the images to the user for perusal in this sorted order. The user then makes a new selection, at which point the database is resorted relative to the new selection. This process repeats until the user finds the sought-after image (or, failing to find it, tires and give up).

One problem with this interface is that the search method it employs is essentially a hill-climbing approach. As such, it is prone to problems with local maxima, and the user can wind up cycling through the same set of faces without making any further progress. It is also not clear how well hill-climbing works in conjunction with a similarity metric such as Eigenfaces that is only roughly correlated with the user's perceptions of similarity (causing the user to sometimes mistakenly guide the search "down" the slope instead of "up"). Another drawback to the Photobook interface is that the user's query is limited to images found readily in the database. This may be especially problematic if the sought-after face is very different from the other database images. An important advantage to the Photobook interaction method is that it uses the natural human ability to recognize faces and thus enables specification of the query without requiring the user to articulate or even be consciously aware of what specific facial features are being sought.

Generally, face recognition systems use image-based queries to solve identification problems. The recognition system typically begins with a digital image of a face to be identified and compares it to images of known individuals in the database. The mug-shot search problem differs from the face recognition problem in that there is no on-line digital image available at the outset to serve as the query. Another important difference is that a mug-shot search system must retrieve faces that look similar to the query face, while a face-recognition system's task is to retrieve other images of the same face. It is not clear that a method that works well for one problem will necessarily work just as well for the other.

Photobook handles the lack of an input query by permitting the user to select query images from the database. An alternative is to enable the user to create or construct query images from scratch. A number of content-based retrieval systems use this approach [QBIC95] [JFS95], but they typically do not provide a creation interface that works well for faces. Furthermore, creating a specific desired face from scratch is a challenging and time-consuming task and it would not make sense to attempt this if suitable database faces are handy.

Systems for producing composite sketches for criminal identification, such as CompuSketch or Identikit, enable a user to create facial images easily, but they typically do not address the database search problem. One such composite sketch system is FacePrints [CJ91], which uses an interactive genetic algorithm [Gol89] to allow a user to create a composite by rating randomly generated "populations" of proposed Identikit-like faces for their similarity to a perpetrator. FacePrints' designers claim that their approach is more effective than traditional systems that require a user to specify individual face parts because it uses a recognition-based rather than an isolated-feature-recall strategy, and is thus better suited to the way people remember faces.

Phantomas, a commercially available automated facial database search system out of Germany, claims to work well with composite sketches as well as photographs as input. However, it does not integrate the creation and search components and the advertised search times (11 minutes for 10,000 images on a Pentium-90 PC [Web98]) do not yet sound practical for interactive search. A study by Hancock, Bruce, and Burton [HBB97] compares the Elastic Graph Matching recognition algorithm [LVBL93] used by Phantomas to several PCA-based approaches and suggests that Elastic Graph Matching may be somewhat better at capturing human perception of similarity between faces.

Recently, several prototype systems that do attempt to integrate composite face creation techniques with database search have been reported. The SpotIt system [BM96] uses eigenfeatures [MP94], applying PCA to pre-annotated facial features, such as the hair, eyes, nose, and mouth. The creation interface produces Eigenface reconstructions from the eigenfeature weights. The user manipulates sliders to select the desired weights for each feature while the system continuously responds to these selections by updating the reconstructed "composite" image. Simultaneously, the system also displays those faces from the database that are most similar to the composite. The weights from an existing database face may be incorporated into the composite. Another system, CAFIIR [WALD94], uses a combination of feature-based PCA weights, facial landmarks, and text descriptions to construct index keys for an image. CAFIIR's composite face creation method permits the user to construct a face from a database of feature parts by blending each part onto a template facial image whose corresponding feature is appropriately warped (using the feature landmark positions) to receive it. CAFIIR permits the user to select one or more of the retrieved images to be used as feed-

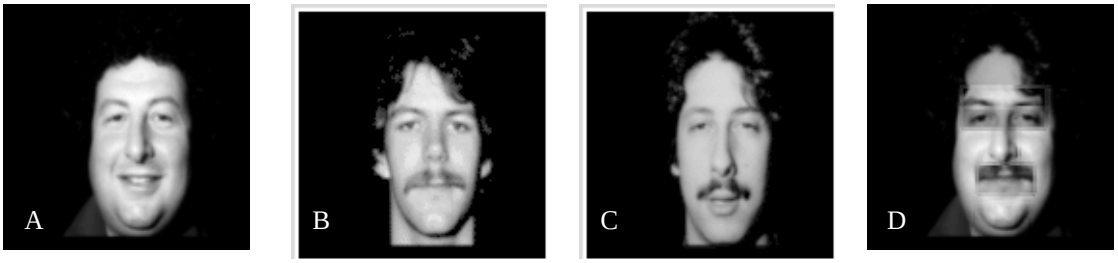

Figure 1. The composite D was created with the cheeks, nose, and chin from A, the mouth and eyebrows from B, and the forehead and eyes from C.

back to refine the search, although these appear not to be used to refine the composite directly. A side benefit to systems like SpotIt and CAFIIR is that, in the event the database search fails (perhaps because the target face is not present), the user is left with a composite of the face that may be used to locate the person via other means.

Photobook, SpotIt, and CAFIIR provide a wide assortment of mechanisms for enabling a user to deal with the "mug-shot search problem." Although the various ideas embodied in these different systems are fascinating, little work has been done to attempt to evaluate their usefulness as applied to mug-shot search, or to try to understand what kinds of search strategies employed with them are most successful. Our goal is to evaluate the benefit of various mechanisms and strategies and to understand how and why any such benefits are obtained.

3 A Prototype System

To conduct our research, we built a simple system that integrates a query image creation method specifically designed for faces with a face-recognition-based retrieval method. Its approach to composite face creation is a hybrid one, using Identikit-like cut-and-paste methods similar to those found in CAFIIR, combined with random composite generation similar to that found in FacePrints (though without the genetic algorithm). For image retrieval, it uses the whole-image based PCA method taken directly from Photobook. (Eigenfeature-based retrieval similar to that found in SpotIt and CAFFIR has also been implemented, but was not used in this study.) The system maintains the original functionality of Photobook, but adds to it the ability to produce composites and to sort the database by distance from them. The creation and recognition subsystems may be used in an integrated fashion, so that interim composites can be used to search the data and interim database search results can, likewise, be used to improve a developing composite.

3.1 The Data

The database we use for testing is a subset of the original Photobook face database. We eliminated most of the multiple images of individual faces, attempting to use the one image with the most neutral expression. Our final test database has approximately 4500 images of faces of varying gender, age, and race. We use the eigenfaces and associated coefficients (weights) as calculated for Photobook [PPS94]. These included 100 eigenfaces produced from a training set of 100 images selected randomly from the database. We use all 100 weights to calculate the Euclidean distance between images. The images consist of 128^2 pixel intensity values and were already eye-aligned as a preprocessing step for calculating the eigenfaces and weights. In addition to the known eye locations, we added annotations for the position of the eyebrows, tip of the nose, center of the mouth, top of the forehead, and bottom of the chin. These annotations were created by hand, though this could be done automatically or semi-automatically using one of several known techniques (e.g., [BP93],[TP91]).

3.2 Composite Creation

Our composites are constructed out of face parts from the images in the database. The feature annotations and eye-alignment made it possible to automatically recombine face parts from several different photographs and still get (most of the time) composites in which the pieces fit together fairly well. Starting with a background image, which determines the cheeks and ears, the remaining face parts are superimposed on this background in rectangles of predefined size (see Figure 1). Rectangle edges are minimally blended with the background. The location annotation of a particular feature is inherited from its source image, so the process of annotating the composites is fully automated. Although we could have allowed the feature locations to move (e.g., placing the mouth lower or the eyebrows higher), as is done in FacePrints, we traded that flexibility for a simpler user interface. The results are generally good, but due to lighting, pose, and feature size variations in the images, some problems do arise. For example, the minimal edge smoothing is not always sufficient to blend the differences when a feature from a very dark face is superimposed on a very light face.

Much of the crudeness that does arise could be eliminated with more sophisticated image blending methods or preprocessing normalization methods, such as those used or proposed in SpotIt and CAFIIR.

The user may tag any number of images from the database as "currently selected." At any time, the user can request a set of random composites to be created from the current selections (i.e., the individual face parts are each chosen randomly from among the current selections). While viewing a set of newly generated composites, the user may choose to add one or more of them to the current selections. These new selections, in turn, are used to generate subsequent composites. As in FacePrints, the system permits the user to fix (and "unfix") individual features when generating random composites so that all random composites will have a particular feature. Manual editing to select an individual feature from one face and paste it onto another is also permitted. By including both manual editing and random composite generation, we hope to obtain the best of both worlds, enabling users to employ both holistic face recognition ability and isolated feature recall ability.

3.3 Eigenfaces Applied to Composites

Since composites are produced from the original database images and inherit all their feature locations from them, the composites maintain the eye alignment and general structure of the originals. The original database images were projected onto the eigenfaces in a preprocessing step, but this operation is fast [TP91], and can be performed on a composite in real time. Thus, we can calculate a composite's weights (i.e., project it onto the eigenfaces to get its location in Eigenface space) on the fly. Once the weights are obtained, the database can be sorted by distance from the composite just as it can any database image. The entire project-and-sort operation is done in response to a single mouse-click. On a 180 MHz Pentium Pro with 64 megabytes of memory, this operation takes under a second for our 4500 image database.

4 The (Pilot) User Study

The user study described here included eleven subjects from our department (students and administrative staff). Its intended focus was on the high-level functional specification of a user-interface rather than the specific implementation details for each function. Nonetheless, implementation details and their associated impact on ease of use can also have a big effect on the success or failure of an interactive system. For example, the specific interaction method used to implement feature editing (e.g., cutting and pasting a nose from one face to another) can have a big impact on how willing a user is to employ that function. Hoping to factor out any possible detrimental effects of our specific implementation choices, for some tasks we allowed subjects to specify their instructions to an expert operator. All subjects worked from the same automated interface that dictated the specific nature and sequence of tasks they were to perform. However, for carrying out feature edits and for recording ranking decisions, they could specify their instructions verbally and by pointing to the screen rather than by directly manipulating the mouse themselves.

Our database could be pre-filtered using text annotations to limit a search to images of the correct gender, age, and race. Since this type of pre-filtering advantage could be applied to all of the approaches we are comparing and would have greatly reduced our database size, we chose not to include it in our experiments.

4.1 Goals

The study is aimed at understanding how best to exploit, in practice, the correlation between the Eigenface and human notions of facial similarity [HBB97]. We strove to assess how well the Eigenface technique works to enable a user to find a face in a database and to learn which search strategies employed with it are most effective. We also strove to determine how much benefit is obtained by adding composite creation to the system.

4.2 The Task

To facilitate analysis of the results, we set up a very constrained set of tasks for all test subjects to complete. In advance, we chose two different target images. Target One, shown on the left in Figure 2, was chosen specifically because the face is quite distinctive. Target Two, shown on the left in Figure 3, was chosen at random. Also in advance, we selected 100 images at random from the database. This same random set was used in experiments for both targets and across all subjects.

Each subject was asked to view Target 1 on the computer screen for several minutes, and was instructed to try to register a clear mental image of the face. It was explained to subjects that they would later be asked to perform tasks that relied on their memory of it (though they were not told what tasks). When the subject was satisfied with the quality of their mental image, the target image was removed from view. Next, each subject was shown the 100 random faces in a kind of computerized mug-book presentation. The screen display fit 20 faces at a time, so there were five sets through which the subject could page back and forth. The subject was asked to select five faces from among these 100 that they felt looked most similar to the target. Selecting five was required even if the subject found this difficult. Once five faces had been selected, the subject was

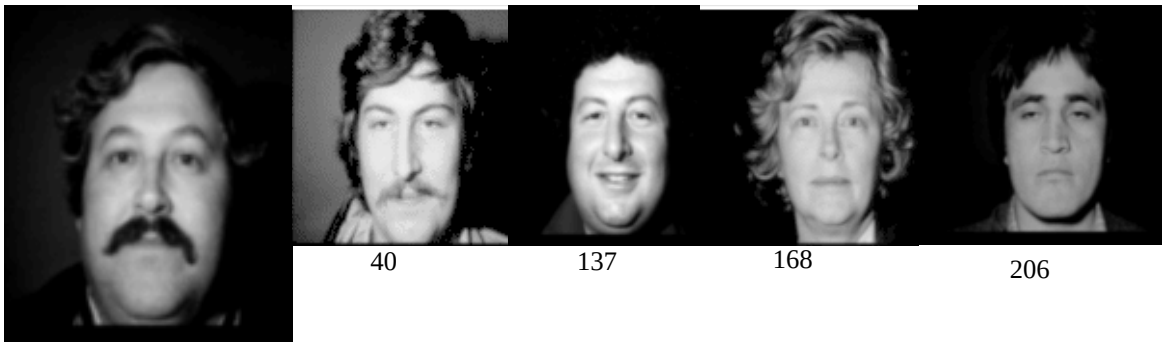

Figure 2. Target 1 and the four faces (out of 100 chosen randomly) closest to it in eigenspace. The number under each image indicates the number of inspections that would be required to find the target using that image as the query.

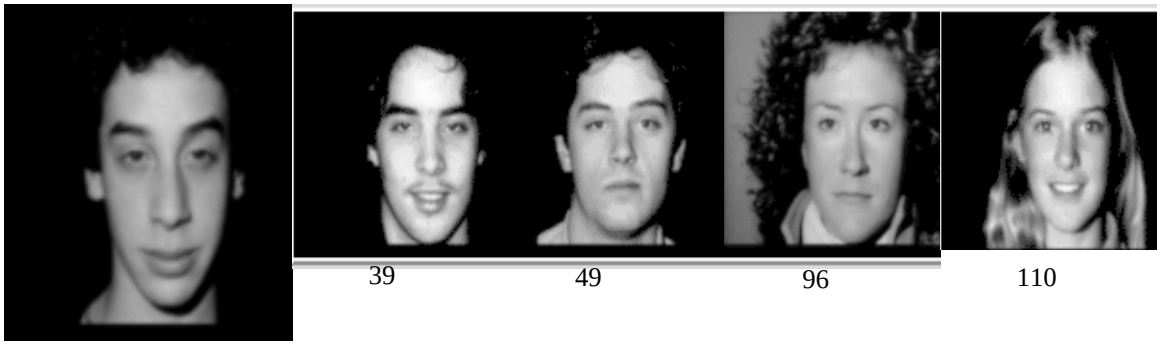

Figure 3. Target 2 and the four faces (out of 100 chosen randomly) closest to it in eigenspace. The number under each image indicates the number of inspections that would be required to find the target using that image as the query.

instructed to rank them for their similarity to the target, from best ("closest") to worst ("furthest"). The subject was permitted to modify the rankings (in an on-screen display of the five images in rank order) until satisfied. The four faces out of the 100 random ones that are actually closest in Eigenface space to Target 1 and Target 2 are shown at the right in Figures 2 and 3. If the human notion of similarity correlated perfectly with the Eigenface distance metric, we would expect these faces to be the user's top four choices. (One might guess from looking at these faces that such perfect correlation does not exist.) Beneath each face is the position (or rank) of the target in the sorted list (of *all* 4500 database images) obtained by using that face as a query image. This number indicates how many image inspections would be required by the user to locate the target face if that image were submitted as a query. We can see from these numbers that selecting the closest image in Eigenface space to use as a query would enable the user to find either target in approximately 40 image inspections plus the initial 100.

After making and ranking the five selections, the system generated and displayed 10 random composites from them (i.e., 10 faces whose parts were selected uniformly at random from among the subject's five selections). The subject was instructed to select one out of these ten random composites that most resembled the target.

Lastly, the subject was asked to attempt to produce a "best" composite via manual editing. The subject could start with either a database image or one of the random composites and modify its features in any way. Subjects could select facial parts from any of the original 100 faces or focus only on parts obtained from their five top choices. Subjects could spend as little or as much time as they wanted producing a final edited composite or on any of the prior tasks. In general, subjects spent between 5 and 45 minutes on the entire set of tasks, averaging about 15 minutes per target.[1] The composite D, shown in Figure 1, is an example of a composite produced for Target 1 by a subject in the study.

When the subject was satisfied with the edited composite, the screen was cleared and Target 2 was displayed. The subject was asked to repeat the same set of tasks for Target 2 as those performed for Target 1. This time, however, the target image remained on the screen for the duration. Hence, for the second target, the sub-

1. Note that, if one could inspect 100 faces a minute, the entire database could be searched in 45 minutes, though this process would likely be extremely tedious and error prone.

ject could work from an on-screen image rather than a mental one.

Allowing the subject to work directly from the target image on-screen rather than from a mental image is potentially problematic because it is a less realistic simulation of the mug-shot search problem. If one actually has an on-screen image of the face sought, then the problem becomes that of face recognition, which is already well-studied and better solved in other ways. Still, allowing the subject to view the target throughout has the advantage that it simulates a photographic memory, thus creating an idealized version of the mug-shot search problem in which differences in visual memory among subjects are factored out of the experiment. This advantage is mitigated somewhat by evidence that people's visual memory of faces plays better to holistic face recognition tasks than it does to isolated feature recall ability. An on-screen image enables a subject to focus on individual features in a way that is less possible when working from visual memory.

We chose to have the subject work from the on-screen image on the second target rather than the first since we expected that by this time in the experiment, work with the first target might have degraded the user's ability to recall a second target. We supposed that permitting the subject to view a target image on-screen throughout the experiment would make it easier to choose the most visually similar images and to produce a good composite. We were interested in comparing results from Target 2 to those from Target 1 where the subject was working from a mental image. Oddly, the composites for Target 1, which were produced from a mental image, were often better (both perceptually closer and closer in Eigenface space) than the composites for Target 2. It is unclear from this small study whether this was a result of the different exposure methods, or simply due to different characteristics of the target faces themselves, or some other factor.

4.3 Evaluation Methodology

We use the *mean number of image inspections* required by the user as a scoring metric for comparisons between strategies. We make the assumption that this metric is more important than the total time required because a user's mental image seems to degrade as more and more images are viewed. We define the *score* of an image, I, (with respect to a target, T) as the position or rank of T in the list of images obtained when the database is sorted by distance from I (this corresponds to the number of image inspections required to find the target if image I is used as a query). We define the *search score* of a *strategy* as the total number of image inspections required to find the target using that strategy. Our database contains approximately 4500 images, so searching it sequentially would, on average, require a user to inspect half the database, resulting in a strategy search score of 2250. We use this as a rough baseline for comparison.

In the study, subjects identified possible query images by picking them from a set of images that were randomly selected from the database. For the purposes of a best-case analysis, we assume that the user can immediately identify the best of these N random selections by picking the one that is most perceptually similar to the target (where "best" is defined as the one with the lowest score). Based on a simplifying assumption,[2] it can be shown that the expected score of the best of N such random selections is the size of the database, D, divided by $(N + 1)$ [BS97]. So, for example, given our database of size 4500, the best of 100 randomly selected images would have an expected score of $4500/(100 + 1)$ or approximately 45. This is consistent with our observations about the two targets and 100 random faces used in our experiments (see Figures 1 and 2). Note that, according to this analysis, the sequential search baseline of 2250 corresponds to the expected score of a single random selection from the database, i.e.,when N = 1. Clearly, the more random selections presented to the user (i.e., the bigger the value of N), the better the expected score of the best one. Of course, the user must inspect the N randomly selected images, too, and these inspections must also be included in the total search score, so there is a point of diminishing returns. Thus, for this approach, the optimal expected total search score is limited by the minimum value of $(D/(N + 1)) + N$. This is approximately $2 \cdot \sqrt{D}$. In our case, the function $(4500/(N + 1)) + N$ has a minimal value when N is 67 (yielding a value of 134).[3] This means that, if the user can successfully pick from among 67 random selections the one closest to the target, that pick can be used to sort the database to obtain a total expected search score of 134. This is our best case expected search score and it is quite good in comparison to our worst case baseline of 2250 for sequential search. The Eigenface method (as applied to the mug shot

2. The simplifying assumption is that the score of image P with respect to image T is equal to the score of image T with respect to image P.

3. Had we noted this when we originally designed the user study, we might have chosen 67 rather than 100 for the number of random images from which the user selects. Fortunately, using N = 100, we still get quite close to this minimum of 134, i.e., 100 + (4500/(101)) is about 145. So our choice was also reasonable. Since the 100 initial image inspections are done by all users, we omit them in the search scores in our tables, so, for our study, the correct optimal expected score to use for comparison is actually 45.

search problem) is based on the presumption that the correlation between the Eigenface and human metrics for determining distance (or similarity) is a strong one. We anticipate that the use of Eigenfaces will permit our subjects to do much better than the sequential search baseline of 2250. But how close can they get to the expected search score of 134 that would result if the human and Eigenface similarity metrics correlated perfectly? Following that, what additional benefit, if any, is derived from adding the use of composite creation to the system?

Answering these questions requires some analysis. Though our test subjects did not actually use the systems' Eigenface sorting mechanism, we apply it in a post-mortem analysis of the raw user data. We sort the database by distance from each of the subject's five database selections, as well as from their first-choice composite, and final edited composite. We then note the position number of the target in each such sorted list (i.e., we note the score of each of these potential query images). From this data, we can compute average search scores across all eleven subjects for various search strategies the users might have employed.[4] For example, we can compare how well our subjects would have done on average had they used only the top choice database image as a query vs. how well they would have done had they used only the final edited composite as a query.

Since we know there is some correlation between the Eigenface similarity metric and the human one [HBB97], we might guess that the closest image in Eigenface space (of the 100) would regularly show up somewhere among the user's top five database choices. If so, the strategy of searching in parallel the sorted lists based on these choices would have an expected search score of 225 (plus the 100 initial inspections).[5] Ideally, we want to be able to compare the optimal average search score among all strategies that use one or more database images to the optimal average search score among all strategies that use one or more of *both* the database images and the composites. Such a comparison would permit us to determine how much benefit, if any, can be derived from the use of composites.

Unfortunately, the huge number of possible search strategies prohibits checking our user data for the average search scores associated with *all* of them. However, a simple characterization of most of the reasonable strategies does permit an exhaustive check of those. We define a "database image only" strategy as a triplet (H, D, I), where H specifies how many of the five database images to use, D specifies how deep to look in the sorted lists for these images before going on to the next list, and I specifies how many such "breadth-first" iterations to perform before returning to look "depth-first" in the first list. For example, the strategy $(3, 40, 2)$ sorts the database for each of the top 3 (out of 5) database images, looks 40 images deep in each of the sorted lists, and then repeats this a 2nd time looking at the next 40 images in each list. Finally, if the target image has still not been found, it goes back to searching the remainder of the first sorted list, and keeps going until the target is found. We assume there is no reason to violate the user's ordering of the five images, so we exclude strategies that use the second image before the first, etc. Likewise, we exclude seemingly random tactics such as looking at image 200 in the first list, then image 46 in the second list, etc. We also assume there is no need to look at all possible values for D. Instead, we look only at multiples of 20 (one screenful of images) for the value of D. (Actually, we use 1, 21, 41, etc., so that pure parallel search [D=1] is included.) Finally, we make the assumption that 1000 is a limit on the search score for a strategy, since any more than that would likely tax a user's patience beyond its limit. If the user gives up well before that, it doesn't matter whether the score is 2000 or 3000, so a search that does not succeed in under 1000 image inspections is simply tabulated as a failure and averaged in with a search score of 1000.

The above definition of a strategy does not yet include composites. To include them, we need only to change the definition into a quintuplet (H, D, I, P_1, P_2), where H now indicates how many of the seven images (the original five, plus the two composites) are used, D, and I are defined as before, and P_1 and P_2 specify the position of the composites in the image set. For example, the strategy $(3, 40, 1, 1, 0)$ places the random and edited composites in positions 1 and 0 respectively, thus bumping the database images down to positions 2 through 6. This strategy searches 40 images deep in each of the three lists associated with the edited composite, the random composite, and the top database image, in that order. If that fails, the search continues in the remainder of the list associated with the edited composite. The set of strategies included in this definition is small enough that we can perform an exhaustive search of all of them, calculating the average search score of each from the raw user data collected in the study.

4. We make the assumption that the user would recognize the target face were it to reappear.

5. Since 45 is the minimum expected score out of the 100 random selections from which the user is picking, we compute $5 \cdot 45$ (to account for the parallel search) to get 225.

TABLE 1.

Strategy	Target 1 average scores	Target 2 average scores
(1, 0, 0, 5, 6) —use top database image	762 (658 with 5 failures)	1238 (729 with 5 failures)
(1, 0, 0, 0, 6) —use top random composite	277 (277 with 0 failures)	1030 (692 with 6 failures)
(1, 0, 0, 6, 0) —use final edited composite	454 (379 with 1 failure)	713 (475 with 3 failures)

TABLE 2.

Optimal Strategies	strategy	average search score	failures
Target 1: Database only	(4, 41, 4)	223	none
Target 1: Database + Composites	(6, 41, 1, 0, 2)	160	none
Target 2: Database only	(5, 61, 2)	577	5
Target 2: Database + Composites	(6, 61, 1, 6, 0)	382	2

5 Results

Recall that our raw data consists of computed image scores (per target) for each subject's top five database choices, top choice random composite, and final edited composite. Due to space considerations, we report here only average scores over all subjects, but the complete raw data from which these averages are calculated is available elsewhere [BS97]. Table 1 above shows the average search scores for three strategies that use only a single query image —either the users' top choice database image, the users' top choice random composite, or the users' final edited composite. The scores outside parentheses show the plain average, whereas inside the parentheses is the average computed with individual failures limited to 1000. From this table, several facts are clear. First, the strategy that uses only the top choice database image does substantially better than the sequential search baseline score of 2250. As anticipated, using Eigenfaces, even in this simple manner, is a substantial win. Still, these scores are a far cry from the expected score of 45 if the users' and Eigenface similarity metrics are perfectly correlated. The second observation we can make is that using the edited composite works much better than using the subject's top database choice. In the case of target 1, using the subject's top choice random composite is even better than using the edited one. Thus, if the user's strategy is constrained to selecting a single query image, using a composite seems like a good idea.

But what about strategies that use multiple query images? Given our more flexible definition of a strategy, how does the optimal strategy using one or more of *both* composites and database images compare to the optimal using one or more of *only* the database images? For each target, we calculated the average search score over all subjects for each possible strategy included in our definition. The first four rows of Table 2 show the optimal strategies (with and without the use of composites) that were identified by this exhaustive search. As we suspected, in the case of both targets, the optimal strategy uses a mix of database images and composites, and is substantially better than the best strategy that uses only database images. Though somewhat similar, the optimal strategy is not exactly the same for both targets. For target 1, the random composite is placed first in the sequence, whereas for target 2, only the edited composite is used. With more extensive user studies it may be possible to determine which strategies are more globally successful.

These results give a clear indication that the use of composites provides a potential advantage over restricting users to database images for their queries. Strategies that include composites seem to enable a user to locate a target face in fewer image inspections and with fewer failure cases. The results also indicate that, for the Eigenface similarity metric, parallel search strategies employing multiple user choices are more effective than strategies that focus only on a user's top choice image, even when that image is a composite produced expressly to look similar to the target.

6 Discussion and Future Work

There are three main avenues for seeking improvement in mug-shot search systems. The first is to attempt to improve the correlation between the human and system metrics for determining similarity between faces. The second is to determine search strategies that best exploit whatever correlation does exist and attempt to build those strategies directly into the system. The third

is to seek a query formulation interface that best facilitates easy construction or location of a query image matching the mental one. There is plenty of potential for improvement in each area, and progress in one area may affect progress (or the need for it) in another.

Our study shows that the Eigenface method, though helpful, is an imperfect model of human perception of similarity between faces. Applying a novel evaluation methodology to our system, we have made progress at identifying specific search strategies that, given the imperfect correlation between the system and human similarity metrics, attempt to use Eigenfaces to the best advantage. We have also shown that the use of facial composites as queries is advantageous compared to restricting users to database images for their queries.

In our study, subjects were limited to a very restricted set of actions within the system. In reality, the system provides a great deal more flexibility than this. At every stage there are many strategy choices to be made. In addition to deciding which images to use as queries and how far down each sorted list to search, the user must decide which, if any, of the images from these sublists should also be used as queries, which images to select for composite creation, how many random composites to generate, whether and when to use manual editing, etc. While sometimes a big benefit, all this freedom can also hinder the user, making the system more complicated and providing many opportunities for costly walks down blind alleys. Our analysis suggests that parallel search strategies using both database images and composites as queries are most successful. With more extensive user studies, we will seek to establish more precisely which strategies are globally successful so that additional "guidance" can be incorporated into the system. We also expect to take a closer look at which features of the query formulation (i.e., composite creation) interface are most useful and at how this affects the tradeoff between simplicity and functionality. In addition, we want to better understand the effectiveness of the interactive refinement approach to building a query image. Does this kind of hill-climbing (i.e., iterating the *select*, *sort*, *search* sequence) really work better than simply selecting one or more images from a random set (as was done in the study described here)? Does hill-climbing suffer from classic problems with local maxima and, if so, does the use of composites help the user get unstuck?

A number of improvements and alternatives to the basic Eigenfaces method have been described in the literature [e.g., MP94, LTC95, WW97, LVBL93]. Most evaluations of these metrics have focused on their success at face recognition rather than similarity retrieval. A number of other general image recognition methods as applied to interactive database search have also been reported [JFS95, RM97], but these have been tested primarily on general image databases rather than specifically with faces. Although it is possible that applying one or more of these methods to the mug-shot search problem will provide improvements over the basic Eigenfaces method, it is not yet clear which method is best. For mug-shot search, the important factor is the strength of the correlation between the human and system metrics for assessing similarity between faces. The best method for this task may be different from the best method for identifying facial images of the same person or for finding similar images in a general image (i.e., non-facial) database. While our study focused on identifying successful strategies and query formulation features in a system employing full face Eigenfaces, for systems that employ other (possibly better) mechanisms for determining similarity between images, the answers may be different. However, the evaluation methodology we describe is a useful tool that can be generally applied to the design and analysis of similarity-based retrieval systems. It can be used both to determine the best search strategies for a given metric and to help distinguish between the many possible candidate metrics.

7 Acknowledgements

The authors would like to thank Alex Pentland, Adolph Baker, and Stuart Shieber (who suggested the characterization of a "strategy" used in our evaluation methodology) for helpful discussions.

8 References

[BM96] R. Brunelli and O. Mich, SpotIt! an Interactive Identikit System, *Graphical Models and Image Processing*, Vol. 58, No. 5, pp. 399-404, September 1996.

[BP93] R. Brunelli and T. Poggio, Face Recognition: Features vs. Templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 10, October 1993.

[BS97] E. Baker and M. Seltzer. The Mug-Shot Search Problem. *Harvard University Center for Research in Computing Technology, Technical Report-20-97*, 1997.

[CJ91] C. Caldwell and V. S. Johnston. Tracking a Criminal Suspect Through Face Space With a Genetic Algorithm. *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 416-421, 1991. Morgan Kaufmann Publishers.

[Gol89] D. E. Goldberg. *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley, 1989.

[GR95] V. N. Guidivada and V. V. Raghavan, Content-Based Image Retrieval Systems. *IEEE Computer*, Vol. 29, No. 9, pp. 18-22, September 1995.

[HBB97] P. J. B. Hancock, V. Bruce, and A. M. Burton, A Comparison of Two Computer-Based Face Identification Systems With Human Perceptions of Faces, Vision Research, in press, 1998.

[JFS95] Charles E. Jacobs, Adam Finkelstein, David H. Salesin. Fast Multiresolution Image Querying. Proceedings of SIGGAPH 95, in Computer Graphics Proceedings, Annual Conference Series, pages 277-286, August 1995.

[KS90] M. Kirby and L. Sirovich, Application of the Karhunen-Loeve procedure for the Characterization of Human Faces, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No. 1, 1990.

[LTC95] A. Lanitis, C.J. Taylor, T.F. Cootes, An Automatic Face Identification System Using Flexible Appearance Models. *Procs. of the 5th British Machine Vision Conference*, vol 1, pp 65-74, ed. Edwin Hancock, BMVA Press, York, UK, 1994.

[LVBL93] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen, Distortion Invariant Object Recognition in the Dynamic Link Architecture, IEEE Transactions on Computers, 42:300-311, 1993.

[MP94] B. Moghaddam, A. Pentland. Face Recognition using View-Based and Modular Eigenspaces. *Automatic Systems for the Identification and Inspection of Humans, SPIE* Vol. 2277, July 1994.

[PPS94] A. Pentland, R.W. Picard, S. Sclaroff. Photobook: Tools for Content-Based Manipulation of Image Databases. *Proceedings SPIE Storage and Retrieval Image and Video Databases II*. Vol. 2,185, 1994.

[QBIC95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by Video and Image Content: The QBIC System, *IEEE Computer*, Vol. 28, No. 9, pp. 23-31, September 1995.

[RM97] S. Ravela and R. Manmatha, Image Retrieval by Appearance, SIGIR 97, Philadelphia, July 1997.

[TP91] M. Turk, A. Pentland. Eigenfaces For Recognition. *Journal of Cognitive Neuroscience*, May 1991.

[WALD94] J.K. Wu, Y. H. Ang, P. Lam, H.H. Loh, and A. Desai Narasimhalu, Inference and Retrieval of Facial Images, *Multimedia Systems*, 2:1-14, 1994.

[Web98] Product Description on the web: http://www.zn.ruhr-uni-bochum.de/work/k1/s110e.htm

[WW97] W. A. S. M. Wahid, Masters Thesis, M. I. T. Media Lab, 1997.