COMPUTER IMAGE RETRIEVAL BY FEATURES: SELECTING THE BEST FACIAL FEATURES FOR SUSPECT IDENTIFICATION SYSTEMS

Eric Lee*
Management Science
St. Mary's University
Halifax, NS, Canada B3H 3C3
elee@bootless.stmarys.ca
(902) 420-5734

Thom Whalen
Division of Behavioural Research
Communications Research Centre
3701 Carling Avenue, Ottawa, Canada
thom@rick.dgbt.doc.ca
(613) 990-4683

Abstract

Correct suspect identification of known offenders by witnesses deteriorates rapidly as more are examined in mugshot albums. Feature approaches, where mugshots are displayed in order of similarity to witnesses' descriptions, increase identification success by reducing this number. System performance depends on selection of system features. Four methods of selecting features are evaluated empirically: theory, random, hill-climbing algorithm, and hybrid. The theory asserts success depends on five properties of system features: informativeness, orthogonality, sufficiency, consistency, and observability. Comparing system performance on the best 10 features selected (from a pool of 90) by each method supports our contention. In four experimental tests of a system with 1000 official mugshots, over 90% of witness searches resulted in photos of target suspects retrieved in the first ten mugshots displayed for examination (using all 90 system features). On average, suspects were retrieved in the first 54, 7, 22, and 70 mugshots when using only the best 10 model features. Hybrid and hill-climbing algorithms did not improve on this performance, and performance of randomly selected sets of 10 features was poor.

KEYWORDS: computer image retrieval, information retrieval, feature retrieval, suspect identification.

[1]Research supported by grants to the first author from the National Science and Engineering Research Council of Canada

CIKM '94- 11/94 Gaithersburg MD USA
© 1994 ACM 0-89791-674-3/94/0011..$3.50

1 INTRODUCTION

Feature approaches to suspect identification, in which witnesses describe facial features of offenders, improve witness accuracy by reducing the number of mugshot photos examined. These computerized systems are intended to replace the ubiquitous mugshot album. Performance of feature systems depends on the set of facial features. We propose, and test, several methods for selecting features.

The key to solving crimes is frequently identification of suspects by witnesses or victims. Police forces typically rely on three methods to identify suspects: verbal descriptions; composite methods such as Photofit or Identikit; and mugshot albums.

Each method has a role to play. However, verbal descriptions lack sufficient detail and accuracy to be of much use [2]. Composite procedures are relatively unsuccessful, because they do not generate accurate, detailed images of suspects [3]. The mugshot album approach is most successful, because people are good at face recognition [4].

Nevertheless, the album approach suffers fundamental problems. The task is tiring, time-consuming, and confusing as witnesses often examine thousands of photos. The probability of selecting the correct suspect (a hit) decreases rapidly after the first 100-200 examined, while the probability of selecting the wrong person (a false alarm) increases rapidly [1, 4, 8, 9, 13].

To improve the album approach, several research teams have developed what we refer to as feature approaches to image retrieval [4, 6, 7, 13, for example]). A set of features is used for distinguishing among mugshot photos in a database. In our feature retrieval system, for example, users describe a suspect (i.e., whose photo they wish to retrieve) in response to a set of 90 queries (e.g., on a 5-point Likert scale, Height: How tall was this person?). The system matches the witness' description with those of mugshots in the database. Mugshots are presented to witnesses in order of similarity to their description.

We empirically tested our system in three experiments [10, 11]. Mugshots of target suspects were typically among

the first 10 examined by subject witnesses for a database of 640 suspects. This performance is significantly better than that for mugshot albums.

However, it was Harmon [7] who, in a classic paper, introduced the feature approach to suspect identification. His approach, while similar to ours in the use of Likert rating scales, required 12 police raters per mugshot. Ours requires one. Harmon's requirement for so many raters per mugshot substantially increases system development costs. In any case, Harmon failed to establish the value of his system empirically. In the only extant empirical test, subject witnesses rated suspects directly from photos, and not from memory.

In contrast, Garner colleagues [5] empirically tested their feature system (witnesses described target suspects from memory). Their system, based on Likert rating scales for some facial features and physical measurement of others (from photos), was superior to the album approach. Subject witnesses identified target suspects more frequently and misidentified incorrect suspects less frequently when using Ellis' feature system. Though successful, Ellis' system requires 10 police raters/mugshot, and mugshot photos must be taken under exceedingly rigorous conditions (faces must be in the same position for replicable physical measurement of features).

Empirical tests suggest feature systems possess valuable qualities. First, they tolerate errors, as a user's description may differ from the database's on several features, and still the target image is retrieved. Second, they permit uncertainty. A user can skip a feature without terminating the retrieval process. Third, they require no instructions or training. The method therefore is well suited to non-expert users.

While system performance has been good, the question remains how to improve it. No consensus exists in literature on the criteria recommended for selecting features. Recommended criteria include saliency [1], reliability and independence [4], usefulness, stability, efficiency, and ease of coding [6]. Features have also been derived from Identikit [13] and the analysis of free descriptions [1]. This lack of consensus, coupled with the complete lack of any theory or empirical evidence justifying the criteria advocated, has hampered development of feature systems.

The purpose of the present paper is sixfold: first, to propose a communication theory for selecting system features; second, to propose several alternative feature selection methods; third, to test these methods empirically; fourth, to assess how system performance changes with number of system features; fifth, to assess feature system tolerance for witness errors and omissions; and sixth, to discuss methodological and theoretical issues arising from these investigations.

The next section presents a theoretical analysis of feature retrieval systems which can be used to guide system development and improve system performance. Four major propositions are advanced: feature retrieval is a communication process; retrieval success is governed by the degree to which five properties of the feature set are met; unifying those five properties is the concept of uncertainty; and, to optimize system performance, those same properties should constitute the criteria for selecting features.

1.1 COMMUNICATION MODEL OF FEATURE RETRIEVAL

The central thesis of this study is feature retrieval is a communication process. The source of a communication is the user, and the destination is the retrieval system. In suspect identification, for example, a witness must communicate to the system an accurate description of a suspect's features to enable the system to retrieve their photo from the database of mugshots. The system assumes the suspect's photo is in the database. However, feature descriptions are transmitted, not actual images. Thus, feature descriptions of each image in the database constitute the set of possible messages. In transmission from user to system the message, or featural description of an image, may be changed inadvertently. Such changes are attributable to addition of unintended information called noise which causes errors and distortions in transmission. (See [4] for a general discussion of information theoretic concepts.)

The effectiveness of feature retrieval systems depends upon their ability to retrieve target images accurately. Retrieval success depends, in turn, upon the ability of the system to discriminate successfully among database images, and the ability to transmit information about target images from user to system.

A major thesis of the present analysis is five characteristics of system features govern the effectiveness of discrimination and transmission: informativeness, orthogonality, consistency, observability, and sufficiency.

1.1.1 Definitions of Factors Governing Retrieval Success

Informativeness is information value (or, equivalently, uncertainty) associated with a feature. Informativeness depends on the number of possible values for a feature and the distribution of database images across the possible coding values. For a scale with v possible coding values, a feature is most informative if it divides the database images into v equal-sized groups. Thus, for a 2-point scale on height, a 50-50 split, such that half the individuals in the database are tall and half are short, is most informative. We measure informativeness in terms of bits.

Orthogonality refers to independence, or lack of correlation, between features. Highly correlated features convey primarily redundant information. Such features convey little new information relevant to discriminating among images. Our measure of orthogonality is Pearson r correlation.

Sufficiency is measured by the power of system features to discriminate among database images relative to the minimum required to uniquely discriminate among them. A sufficient number of features is necessary to ensure adequate discrimination and accurate retrieval.

Consistency is the degree of agreement between database and the user codings. Any deviation between a user's coding of a feature and that by the database coder causes an error (in transmission). We estimate consistency by the mean deviation between user and database feature ratings, or by the percentage of deviations greater than x scale units.

Observability is the degree to which a given feature is actually used by system users for retrieval purposes. For example, in suspect identification, witnesses are less likely to observe or recall ear lobe size than hair colour. We estimate observability by the percentage of features omitted by users.

A simple example will illustrate these concepts. In suspect identification, suspects are described using 5-point rating scales on facial features such as hair length and size of nose. Our theory asserts hair length is a good feature to the extent that there are an equal number of suspects in the database coded by police as 1's, 2's, ..., 5's (i.e., for 1000 mugshots, 200 1's, 200 2's, ...); witnesses do not omit describing hair length frequently or make large errors (e.g.,

police rate target suspect as hair length as 5 but witness rates same person as a 2); and provided it is uncorrelated with nose size and other system features.

The analysis which follows serves to justify claims that these five factors govern effectiveness of discrimination and transmission of a set of features, and that uncertainty, which unifies them, is the mechanism by which they influence system performance.

1.1.2 Basic Information Concepts

The questions to be answered in communication systems are centered around the amount of information transmitted; the coding process used to change a message into a signal (i.e., from image to feature description); and the effects of noise. The analysis to follow addresses these questions quantitatively. We begin with definitions of some basic information theoretic concepts.

The system's task is to decide which image in the database matches the user's description. To begin, any of the database image descriptions might match the user's description of the target image, and system uncertainty is at a maximum. Information is communicated only to the extent that this original array of possible descriptions is reduced. Information and uncertainty are intimately related. If we can measure uncertainty, then we can measure information as reduction in uncertainty. Uncertainty is quantitatively related to the number of database images. If the system is told nothing about the target, then all database images are possible matches, and uncertainty is maximum.

The amount of information conveyed by a given feature is a measure of how much it reduces the array of possible images. If, for example, half the suspects in a mugshot database are tall while the others are short, then knowing a suspect is tall reduces the array of potential suspects by one-half. Every time the number of possible images is reduced in half, one bit of information is gained.

Together, two features can reduce uncertainty by an amount referred to as joint uncertainty. Uncertainty associated with a second feature may overlap that associated with a first feature. Thus, two features may redundantly convey some of the same information. Contingent uncertainty is the average uncertainty in common between two features (uncertainty associated with the intersection of two features).

These concepts provide the foundation for our theoretical analysis of feature-matching retrieval. The analysis is divided into two sections: discrimination and transmission. Discrimination depends upon informativeness, orthogonality, and sufficiency, whereas transmission depends upon consistency and observability.

1.1.3 Discrimination: Uncertainty Associated with the Database

Theoretically, a given set of d features can reduce uncertainty by a maximum given by the nominal uncertainty of the feature set. Nominal uncertainty, U_{NOM}, equals the sum of the maximum possible univariate uncertainties of each feature.

Correlated features and unequal distributions introduce redundancy. Redundancy reduces the amount of uncertainty reduction which a given feature set can achieve. Nominal uncertainty associated with a given set of features can be partitioned into three independent, non-overlapping components: actual uncertainty associated with the feature set U_{FS}, uncertainty associated with unequal distributions U_{DIST}, and uncertainty associated with correlated features U_{COR}. In general, the actual amount of information conveyed by a set of d features U_{FS} will be less than, or equal to, U_{NOM}.

Information in the feature set may be insufficient to permit full discrimination among all database images. Full discrimination requires $log_2 n$ bits of information (which is called database uncertainty, U_{DB}). Sufficiency of a given feature set is reflected by the ratio of uncertainty in the feature set to actual database uncertainty. If every image in the database is to be correctly distinguished from all others in the database, then uncertainty in the feature set must equal or exceed U_{DB}. For a database containing eight images, $U_{DB} = 3$ bits. Thus, the featural description must deliver at least three bits of information to permit completely accuratediscrimination.

Discrimination depends on the number of system features, that is, on sufficiency. From this analysis it follows that elimination of any system feature necessarily reduces, or at best leaves unchanged, the ability to discriminate. Conversely, adding a system feature can never decrease the ability to discriminate.

Accurate discrimination among database images does not, however, guarantee successful communication of this information from user to system. This information must also be transmitted successfully from witness to system if the system is to accurately retrieve target suspects from the database.

1.1.4 Transmission: Uncertainty of Transmission From User To System

Some information may be lost in transmission from user to system (called equivocation in information theory), while other information may be distorted (referred to as noise). Communication may fail, either because the user fails to observe the same image features as the database coder, or because user and database coders lack consistency (i.e., fail to agree on how to code a given feature). Thus, witness errors and omissions reduce the amount of information transmitted from user to system. Consequently, system performance deteriorates as error and omission rates increase.

The total or joint uncertainty U_{JOINT} includes two overlapping sources of uncertainty: uncertainty in the system feature set U_{FS}, and uncertainty in the user U_{USER}. This total uncertainty may be partitioned into three non-overlapping components: lost information U_{LOST} (i.e., uncertainty residing only in the database); noise U_{NOISE} (i.e., information only in the user); and transmitted information U_{TRANS} (i.e., information common to both database and user) which is the successfully transmitted information essential for correctly identifying target images. We are primarily concerned with maximizing the amount of transmitted information.

1.2 SYSTEM DESIGN AND TESTING

Other researchers, such as Ellis, have defined success somewhat restrictively as successful identification (e.g., hits, misses). Such a position can limit the process of design and development. Our approach is predicated on a wider definition of success including identification success, retrieval rank, system tolerance, and feature quality.

Identification performance is the ability of witnesses to identify suspects successfully from their photos and includes such measures as hits, misses, false alarms, and nonretrievals. Retrieval rank is the rank order of a suspect's photo when the database photos are arranged in order of

similarity to a witness' description. Tolerance performance measures the ability of a system to tolerate witness errors on some features (error tolerance), and omission of others (omission tolerance), and still perform acceptably (as measured by retrieval rank). The fourth class of measures, which we call uncertainty measures, assess informational properties of the features themselves: informativeness, orthogonality, sufficiency, consistency, and observability.

We argue that system performance is most accurately measured by retrieval rank. Earlier presentation of target photos markedly improves identification performance [1, 4, 8, 9, 13, 14]. System and user performance, on the other hand, are confounded in measures of identification success. Therefore, our immediate objective is minimization of retrieval rank.

1.2.1 The Mugshot Database

The database consists of 1000 official mugshot photos of known offenders. (In contrast, Harmon and Ellis used photos of non-offenders.) Colour photos were taken under standard conditions – frontal view of face from the shoulder up (90 x 125 mm prints). The suspects are all white males, aged 18-33 (99.5% are aged 18 to 27).

Each mugshot was coded on 90 facial features by one of 13 raters (6 males and 7 females in their early twenties). The raters received no training or instructions. Raters coded directly from the photo which was always available for inspection, as would be the case if police officers coded the mugshots. Coding time per mugshot was approximately five minutes. Each feature is coded on a 5-point Likert scale (e.g., narrow nose 1 2 3 4 5 broad nose).

1.2.2 The Feature Retrieval System

Our system, programmed in C, is implemented on a Sun Sparc 10. Witnesses work directly on the system. They describe suspects using the same 90 5-point scales. Feature queries appear on the screen in succession.

Similarity between witness and database descriptions is measured by a Euclidean metric as Harmon [7] did, that is, we sum the squared deviations between witness and database feature descriptions and take the square root of the sum. In the Ellis system, similarity is measured by the number of feature matches. Preliminary research in our lab suggests the Euclidean metric minimizes retrieval rank relative to other metrics.

1.3 FEATURE SELECTION METHODS

We selected the best 10 system features (from the 90) in four ways: model, hill-climber, hybrid, and random. These methods are described next.

1.3.1 Model

Using data gathered in previous experiments [10, 11], we selected the 10 most informative, orthogonal, consistent and observable features (selected features high on all 4 criteria).

1.3.2 Hill-climber

We developed a hill-climbing algorithm to improve upon the model as a selection method. The algorithm started with the first feature and computed the retrieval rank for 60 witness searches (the 60 searches derived from Experiments 1 to 3 described below) using only a single feature. This process was repeated for each feature. The singleton feature yielding the lowest mean retrieval rank was retained and a second feature was added to form a seed for the next step. The algorithm proceeded to change one feature at a time in this seed, reassessing each time system performance on the corpus of 60 witness searches. If mean retrieval rank decreased below that for the seed, then the new set of n features became the seed, and the process was repeated. The process was terminated when the best 10 features were identified. The algorithm produces a local minimum but does not guarantee the global minimum. (Exhaustive search of all possible combinations of 90 choose 10 features is not feasible in reasonable time.)

1.3.3 Hybrid

The algorithm, starting with the best 10 model features as a seed, proceeded to change one feature at a time, reassessing each time system performance on the corpus of 60 witness searches If mean retrieval rank decreased below that for the seed, then the new set of 10 features became the seed, and the process was repeated. The hybrid and hill-climber algorithm produced the same set of 10 optimal features. For this reason, the hybrid is not distinguished from the hill-climber in analyses and discussions.

1.3.4 Random

System performance of randomly selected features provides a baseline for other methods of selecting features. We randomly selected 100 sets of 10 features (from the 90 system features). For each set we determined mean retrieval rank for the 60 witness searches derived from Experiments 1 to 3 described in the next section. We chose the best set from among these 100 sets of 10 randomly selected features.

1.3.5 Baseline

For comparison purposes, we report the average of all 100 sets of 10 randomly selected features (average includes the best 10 random features and 99 other sets of 10). System performance on the three feature selection methods just described should be better than random performance. See Table 2.

2 EVALUATION OF FEATURE SELECTION METHODS

The data from three previous experimental tests of our system were reanalyzed to evaluate our feature selection methods [for details see 10, 11]. The database employed in [10, 11] was expanded from 640 to 1000 mugshots for present purposes. Feature descriptions provided by subject witnesses in each experiment were reanalyzed using the larger database. The objectives and methodology originally employed in each experiment are sketched first. Results are then reported for each reanalysis using all 90 system features in assessing system performance. Finally, we report performance on the best 10 system features for each method of selecting features. (Only results pertinent to the present thesis are discussed here.)

2.1 Database Expansion

The database was expanded to include 1000 official mugshot photos by adding 360 to the original 640 tested in [10, 11]. The new photos did not differ from the old in any systematic

Exp.#	Rank
1	24.3
2	6.5
3	3.8
4	7.2

Table 1: Mean Retrieval Rank of Target Suspects When System Uses All 90 Features

way. (See [10, 11] for a detailed description of photos and offender population.) The 1000 mugshots were coded by a total of 13 people, 6 male and 7 female, ranging in age from 20 to 32 (mean = 22.4). All were either students or workers at St. Mary's University. Each mugshot was coded by one of the raters on 90 facial features. The raters received no training or special instructions.

2.2 Experiment 1 Reanalysis

A first empirical test of our prototype was conducted to test its effectiveness in identifying suspects. Subjects entered their descriptions of suspects on our computer system but did not search through the actual mugshots. System performance was measured by retrieval rank of the target rather than identification success. This procedure simplifies the experimental process, reduces time required to conduct the experiment, and reduces the security risk created when subjects view many mugshots.

2.2.1 Method

Five subjects, including males and females in their early twenties, were tested. Subjects first searched for a practice mugshot. Then five target mugshots, randomly selected from the database, were presented, one at a time. Order of presentation was counterbalanced. Subjects had 10 sec to examine a mugshot. (Ellis' testing procedure was followed to increase comparability of our systems.) For each target, subjects answered the 90 feature queries on the computer. They were encouraged to answer all queries, guessing if necessary, though skipping was permitted. No feedback was provided between trials, and subjects were limited to one search per target.

2.2.2 Results

Mean retrieval rank for the 25 witness searches was 24.3 (using all 90 system features). See Table 1.

Table 2 presents system performance results for the best 10 features using each method of selecting features. The methods differed significantly in retrieval rank of target suspects, $F(2,8) = 18.51, p < .01$. All methods differed significantly (except hill-climber and hybrid which were identical since they produced the same set of optimal features) by the Newman-Keuls multiple-comparison test ($p < .05$). Hybrid/hill-climber performance was superior to model performance which was, in turn, superior to random performance.

2.3 Experiment 2 Reanalysis

The primary objective was to test an alternative procedure for eliciting suspect descriptions from witnesses which would discourage indiscriminate guessing and improve recall. Instructions to subject witnesses were changed in two ways. First, they wrote out a suspect's description by listing clearly recalled characteristics. Then they answered our feature queries on the computer. Second, when using the computer system, subjects were encouraged to skip features which could not be recalled with confidence. They were not limited, however, to describing features written down earlier. We expected performance, with prompting, to be superior to that without.

	Baseline	Selection Method			
Exp#	Random	Model	Hill	Hybrid	Random
1	225.0	53.6	21.2	21.2	86.4
2	172.7	7.0	7.5	7.5	52.6
3	152.9	22.0	10.8	10.8	54.4
4	311.0	70.0	139.0	139.0	225.8

Table 2: Mean Retrieval Rank of Target Suspects When System Uses Only the Best 10 Features (For Each of Three Methods of Selecting Features)

2.3.1 Method

Five subjects, aged 21 to 43, were tested, including both males and females. The Experiment 1 procedure was used. However, guessing was discouraged and subjects wrote out a description of a suspect before using the computer system. No practice was given.

Five mugshots, randomly selected from the database, were presented to each subject in counterbalanced order. Subjects had 10 sec to examine each target.

2.3.2 Results

Mean retrieval rank for the 25 witness searches was 6.5 (that is, photos of target offenders were, on average, the sixth or seventh of 1000 database mugshots displayed to witnesses). See Table 1.

Table 2 compares system performance for the three selection methods. The three methods differed significantly in mean retrieval rank, $F(2,8) = 18.28, p < .01$. Model and hybrid/hill-climber selection methods did not differ significantly, but both were significantly better than random (Newman-Keuls test, $p < .05$).

2.4 Experiment 3 Reanalysis

The purpose of this experiment was to use a more realistic testing procedure than the first two experiments (witnesses saw only a single target face, rather than five); to eliminate asymmetric transfer and range effects; to increase generalizability by testing more subject witnesses, and to explore the effect of using witness judgments of absolute certainty on some features to eliminate mugshots.

In the first two experiments, subject witnesses searched for several target offenders. Such repeated-measures designs are plagued by practice, order, carry-over, asymmetric transfer, range and context effects [15]. Analysis of the first two experiments showed no effect of practice on performance. However, the effects of asymmetric transfer and range cannot be eliminated by randomization, counterbalancing, or taking out order as a factor in the design. To eliminate these potential problems and to increase realism

in the present study, subjects searched for a single target offender.

2.4.1 Method

Ten subjects were tested, five male and five female. Ages ranged from 19 to 30. Subjects included university students and secretaries as well as workers from the local community. The procedure was identical to that for Experiment 1 with four exceptions. First, subject witnesses provided a written description of their suspect. Second, subjects were asked to list those features for which they were absolutely certain of their judgment. Only then did they answer the feature queries on the computer retrieval system. Third, after answering the system queries, subjects searched the database mugshots, one by one, in rank order. They were asked to identify the target suspect. Fourth, only a single target was presented to each subject witness. No practice was given. Guessing was not discouraged. Subjects were randomly assigned one of two targets, randomly selected from the database.

2.4.2 Results

Mean retrieval rank for the 10 witness searches was 3.8 (using all 90 system features). See Table 1.

Table 2 presents system performance for the best 10 features selected by each method. The methods differed significantly, $F(2,16) = 11.84, p < .01$. Model and hybrid/hill-climber selection methods did not differ significantly, but both were significantly better than random (Newman-Keuls test, $p < .05$).

2.5 Experiment 4 Cross-Validation

Three of the selection methods – hill-climber, hybrid and random – relied on data obtained in Experiments 1 to 3 to generate the best 10 system features. Such a procedure suggests the mean retrieval ranks reported for these methods might be artificially inflated for those three experiments (see Table 2). In contrast, the model used data from a previous unpublished study (in which six people rated 25 men on the facial features), and not data from those three experiments.

To control for this potential problem, a fourth experiment was conducted to cross-validate the results.

2.5.1 Method

Ten subjects were tested, five men and five women, ranging in age from 20 to 44. Ten mugshots were randomly selected as targets. The procedure was the same as that for Experiment 3 (each subject witness examined a single target suspect's photo for 10 sec).

2.5.2 Results

Mean retrieval rank for the 10 witness searches was 7.2. See Table 1.

Mean retrieval rank is reported in Table 2 for the three selection methods. The effect of method was significant, $F(2,18) = 3.83, p < .05$. The model was significantly better than random (by Newman-Keuls test, $p < .05$). No other pairwise differences between means were significant.

3 DISCUSSION

The system has the properties we want. Retrieval performance is good. In four experimental tests of the system on a database of 1000 mugshots, retrieval rank of target mugshots averaged 12.6 (using all 90 system features). Over 90% of all searches resulted in retrieval ranks less than 10. In Experiment 3, all suspects were successfully identified. Unlike Harmon and Ellis, these results were obtained using photos of known offenders.

This level of system performance is not attributable to short retention intervals and the use of photos as targets. In a recent series of experiments using people as targets, rather than photos, system performance was comparable. In another series of experiments, we tested the effect of time delays up to 3 hours between viewing of suspects and retrieval and found no effect on system performance.

Computer simulations suggest the current system should perform well even when database size is increased to over 10,000 mugshots. Moreover, the system has a high degree of tolerance both for witness errors and omissions. Performance was good even with error rates as high as 25% and omission rates up to 50%.

Four methods of selecting system features were compared: communication model, hill-climbing algorithm, random, and hybrid (hill-climbing algorithm with best model features as seed). Each method was used to select the 10 features most likely to optimize system performance. All four experiments provided independent tests of the model method of selecting features because the best 10 model features were selected using an independent set of data. In contrast, the other three methods all relied on data derived from the first three experiments and are, therefore, not independent tests (each method takes advantage of chance by examining many alternative sets of 10 features and selecting the best of them). Only the fourth experiment, the cross-validation, provides an independent test of these methods.

In four independent tests, the model performed well. Mean retrieval rank across four experiments averaged 34.8 for the best 10 model features (for a database of 1000 mugshots). In all experiments, the model was significantly better than the random method of selecting features. Conversely, in the one independent test of the random, hill-climber, and hybrid methods, the three methods did not differ significantly.

These results support our contention that informativeness, orthogonality, consistency, and observability are important determinants of feature system performance, and consequently, qualify as suitable criteria for selecting system features.

The present experiments and simulations cannot be considered definitive. Nevertheless, the results strongly support our contention that feature retrieval systems hold considerable promise for the future.

Ellis has argued that Harmon's procedure, averaging subjective rating of 10 judges to reduce coding error, effectively precludes the practical use of the subjective approach to feature retrieval. The successful performance of our system supports our contention that multiple judges may not be required for subjectively-based systems.

The present results also argue that the success of our system is not strictly attributable to use of a large set of system features. With just 10 system features (selected by model), retrieval rank averaged less than 70 in each of the four experiments.

References

[1] Davies, G.M., Shepherd, J.W. and Ellis, H.D. Effects of interpolated mugshot exposure on accuracy of eyewitness identification. *Journal of Applied Psychology*, 64(1979), 232-237.

[2] Ellis, H.D. Practical aspects of facial memory. In *Eyewitness testimony: Psychological perspectives*, G.L. Wells and E.F. Loftus, Eds. Cambridge University Press, New York, 1984.

[3] Ellis, H.D., Davies, G.M. and Shepherd, J.W. A critical examination of the Photofit system for recalling faces. *Ergonomics*, 21(1978), 297-307.

[4] Ellis, H.D., Shepherd, J.W., Shepherd, J., Klin, R.H. and Davies, G.M. Identification from a computer-driven retrieval system compared with a traditional mugshot album: A new tool for police investigations, *Ergonomics*, 32(1989), 167-177.

[5] Garner,W.R. *Uncertainty and Structure as Psychological Concepts*. Krieger Publishing, Huntington, N.Y. (1975).

[6] Goldstein, A.J., Harmon, L.D. and Lesk, A. Identification of human faces. In *Proceedings of the IEEE*, 59(1971), 748-760.

[7] Harmon, L.D. The recognition of faces. *Scientific American*, 229(1973), 70-82.

[8] Laughery, K.R., Alexander, J.F. and Lane, A.B. Recognition of human faces: Effects of target exposure time, target position, pose position, and type of photograph. *Journal of Applied Psychology*, 51(1971), 477-483.

[9] Laughery, K.R., Fessler, P.K., Lenorovitz, D.R. and Yorlick, D.A. Time delay and similarity effects in facial recognition. *Journal of Applied Psychology*, 59(1974), 490-496.

[10] Lee, E.S. and Whalen, T. Computerzied feature retrieval of images: Suspect identification, *Ergonomics*, 1993a (in press).

[11] Lee, E.S. and Whalen, T. Computer image retrieval by features: Suspect identification, *Proceedings of the 1993 Conference on Human Factors in Computing Systems INTERCHI'93* at Amsterdam, The Netherlands, (ACM, New York) 1993b, 494-499.

[12] Lee, E.S. and Whalen, T. Feature approaches to suspect identification: The effect of multiple raters on system performance. *Ergonomics*, 1994, (accepted).

[13] Lenorovitz, D.R. and Laughery, K.R. A witness--computer interactive system for searching mug files. In *Eyewitness testimony: A psychological perspective*, G.Wells and E. Loftus, Eds. Cambridge University Press, New York, 1984.

[14] MacGregor, J.N., Lee, E.S. and Whalen, T. The feature matching approach to the computer retrievalof graphics: An enhancement, *Information Services and Use*, 9(1989), 127-137.

[15] Poulton, E. C. Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin*, 80(1973), 113-121.