SUSPECT IDENTIFICATION: TRADITIONAL MUGSHOT ALBUM VERSUS COMPUTERIZED FEATURE SYSTEM

Eric Lee[1], Thom Whalen[2], Chandra Bisesar[1], and Glenda Reid[1]

[1]Management Science
St. Mary's University
Halifax, NS, Canada
elee@bootless.stmarys.ca

[2]Division of Behavioural Research
Communications Research Centre
3701 Carling Avenue
Ottawa, Canada
thom@debra.dgbt.doc.ca

KEY WORDS Computerized image retrieval, computerized suspect identification.

ABSTRACT In suspect identification, witnesses examine photos of known offenders in mugshot albums. Identification success deteriorates rapidly, however, as the number examined increases. Feature approaches, where mugshots are displayed in order of similarity to witness descriptions of suspects, increase identification success by reducing the number examined. In this study, subject witnesses searched for target suspects in a database of 1000 mugshots. Feature system users correctly identified more target suspects (90%) than did album users (60%) and misidentified fewer innocent suspects (0% versus 38%). For album users, identification success declined as the number of photos examined by witnesses increased. For feature users, the photo of target suspects was, on average, the 16th of 1000 photos examined.

1. INTRODUCTION

The identification of suspects by witnesses is important in solving many crimes. Of the methods currently used by police to aid witnesses, the most successful is arguably the mugshot album approach (Ellis et al. 1989; Laughery and Wogalter 1989). Witnesses search through albums of facial photos of previous offenders for suspects. Albums work well when witnesses search through a small number of photos. Identification success deteriorates rapidly, however, as witnesses examine more mugshots (Deffenbacher et al. 1981; Laughery et al. 1974; Lenorovitz and Laughery 1984; Davies et al. 1979). This is one of the most consistent, reliable effects in suspect identification research and is often reported anecdotally by police (Laughery and Wogalter 1989).

To eliminate the weaknesses while retaining the strengths of the album approach, researchers have proposed what we call feature approaches to suspect identification. Feature approaches, in which mugshots are displayed to witnesses in order of similarity to witness' descriptions of suspects' facial features, increase identification success by reducing the number of photos examined. They are like albums in requiring witnesses to search through a series of offender mugshots. They differ, however, in the sequencing of mugshots. In albums, mugshots are commonly unorganized or organized by date of arrest. Occasionally, photos are separated by race, age, and sex, but even then photos are unorganized within a grouping. Consequently, witnesses examine the faces of many men who do not resemble the suspect at all. Feature approaches reduce the number of intervening, dissimilar faces witnesses must examine, thereby reducing confusion and increasing identification success.

1.1. Our Computerized Feature System

In our feature system, raters (only one or two raters/mugshot required) working for police describe offenders from their mugshot on a set of 107 facial features using 5-point scales (see Table 1). Witnesses describe suspects on the same 107 features. Witness descriptions are matched with descriptions of each offender in police files. Our

K. Nordby et al. (eds.), *Human-Computer Interaction*

system rank orders all database photos in sequence from most to least similar to a witness's description.

1.2. Other Computerized Feature Systems

Ours is not the only feature approach to suspect identification. In a classic paper, Harmon (1973) introduced the feature approach to suspect identification. In his system, suspects were coded by 10 raters/photo on 21 facial features using 2- or 5-point Likert rating scales. Harmon failed, however, to demonstrate the viability of his feature system, and it was not developed further.

Ellis et al. (1989) were the first to establish empirically the value of a feature system over the traditional album system. The Ellis system is a hybrid in which most of their 47 facial features are coded using physical measurements taken from the photos. Others are coded subjectively (by 12 raters/photo using subjective rating scales). Witnesses necessarily rate suspects on all facial features using subjective Likert scales.

The major difference among our three feature approaches is the way database mugshots are coded: physical measurement or subjective rating. In subjective rating systems, such as our own or Harmon's (1973), one or more police coders rate suspects' mugshots in a database on subjective scales (e.g., 5-point Likert scale for hair length: short 1 2 3 4 5 long). Our system relies on judgments by at most two or three raters/mugshot, whereas in their systems, judgments of 10-12 raters/mugshot are averaged.

Ellis asserts the main problem with rating approaches such as ours "lies in the development of a database in which every face might have to be described by a group of judges and their responses averaged. This could be prohibitively time consuming as well as error prone." The implicit assumption here is that subjective judgments vary considerably from person to person. Therefore, so the argument goes, multiple database raters are required to reduce this error by averaging their ratings. This view is so widely and strongly held by researchers that it has assumed the status of law.

We disagree. We argue it is a myth, untested and unchallenged. We argue that the effect of variation among raters is certainly to introduce unwanted error. The question is whether this error variation is sufficient to impair retrieval performance. It is not if these errors tend to be small relative to true differences between suspects. The successful performance of our system in previous studies, in which only one or two police raters/mugshot were used, supports our contention that large numbers of judges may not be required for subjective systems (Lee and Whalen 1993; Lee et al. 1993).

1.3. Empirical Tests of Feature Systems

In a series of experimental tests of our feature system, which is based on subjective ratings, retrieval rank of photos of target suspects averages about 20 for databases of 1000 offender mugshots (Lee and Whalen 1993, 1994a, 1994b; Whalen et al. 1994). Thus, witnesses typically search through just 20 out of 1000 photos before examining the photo of a target suspect for systems with 2 raters/mugshot. For single-rater/mugshot systems, retrieval rank averages between 30 and 90. More raters do not improve system performance further. These results hold for both photo and live target suspects. Furthermore, delays of up to three days between viewing and describing suspects have no effect on system performance.

Missing, however, from empirical tests of our system is a direct comparison between feature and album systems. The only direct comparison between feature and album systems was conducted by Ellis et al. (1989). They empirically compared their computerized feature system with a traditional mugshot album search. For typical faces, identification success was significantly higher for their feature system (69% hit rate for feature versus 44% for albums). Feature system users also made fewer incorrect identifications for typical faces (9% false positives versus 47%). Album and feature systems did not differ for distinctive faces (hit rates were very high: 75% and 78%).

While Ellis has established the value of physical measurement feature systems over album systems, the effectiveness of subjective rating feature systems, such as ours, over album systems has not been demonstrated. We propose doing so.

2. METHOD

2.1. The Mugshot Database

The database consists of 1000 official mugshot photos of known offenders. (In contrast, 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 (over 99.5% were between the ages of 18 and 27).

Each mugshot was coded on 107 facial features by one of 13 raters (males and females in their early twenties). The raters received no training or instructions. Our rationale for avoiding such instructions is any operational system must be robust to work well in the field (where the degree of control one can exercise over such matters may be limited). 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).

2.2. The Feature Retrieval System

Witnesses describe suspects using the same 107 item paper-and-pencil questionnaire. They are encouraged to skip any features not remembered clearly. Guessing is actively discouraged. Our system rank orders all database mugshots in terms of their similarity to a witness' feature description of the suspect. Similarity between witness and database descriptions is measured by a Euclidean metric, 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 (Lee et al. 1993).

2.3. Target Suspects

To serve as target suspects for subject witnesses to recall, 50 photos were randomly selected from the database. Consequently, all targets were young, white males. For the feature condition, photos of 10 suspects served as target suspects for subject witnesses. Presentation position of the photos of target suspects was determined independently for each witness by our retrieval system. For the album conditions, ten photos of target suspects were randomly selected from the first 100, the second 100, the third 100, or the fifth 100 database photos (means approximately = 50, 150, 250, and 450). Database photos were presented in the same sequential order for all album subjects. (Order was based on date of booking.)

2.4. Raters

Three "police" raters, two women and one man, described each suspect on the 107 facial features. Raters, all white North Americans, ranged in age from 20 to 45.

2.5. Subject Witnesses

The 50 subject witnesses included 27 men and 23 women ranging in age from 16 to 53 (mean = 28.8 years old). Half the subjects were from business or government, and half were students. Of the 33 subjects describing their race, 19 were white and 14 non-white (e.g., Chinese, East Indian, Indonesian).

2.6. Procedure

Subject witnesses were randomly assigned to one of five experimental conditions: one feature and four album conditions. Each subject witness was instructed to remember the face of a single target suspect. Photos of target suspects were displayed for 10 sec. Album subjects then searched the database mugshots, one photo at a time, in sequential order. Feature subjects first filled out the questionnaire describing the features of the target suspect and then examined the database mugshots in order of similarity to their feature description of the target suspect. No instructions were given on how to use the feature description questionnaire. No feedback was provided until the experimental session was completed.

3. RESULTS

3.1. Comparison of Feature and Album Performance

Identification performance of subject witnesses can be classified into one of three types: correct identification of target suspect (a hit), identification of the wrong person as the target suspect (a false alarm), and failure to identify anyone as the target (a miss). Identification results are displayed in Table 2.

The difference in hit rates between feature (90%) and album (60%) was marginally significant, $\chi^2(1) = 3.20$, $p < .08$. (Significance values are unchanged if Fisher exact probability tests are used.) An a priori 2 x 5 χ^2 analysis of differences among the five conditions (4 album and 1 feature) indicated a significant difference, $\chi^2(4) = 9.48$, $p < .05$. The hit rate for the feature condition (90%) was significantly higher than that for the 450 album condition (30%), post hoc multiple-comparison $\chi^2(1) = 7.50$, $\underline{p} < .01$. The feature condition did not differ significantly from the other album conditions.

Album users were more likely than feature users to identify the wrong person as the suspect, 38% versus 0% false alarms, $\chi^2(1) = 5.36$, $\underline{p} < .02$. An a priori 2 x 5 χ^2 analysis of the differences between experimental conditions (4 album and 1 feature) indicated a significant effect, $\chi^2(4) = 12.38$, $\underline{p} < .02$. The false alarm rate was higher for the 450 album condition (70%) than for the feature condition (0%), multiple comparison $\chi^2(1) = 7.50$, $\underline{p} < .01$.

There was an effect of experimental condition (1 feature and 4 album conditions) on the number of photos examined per search, independent groups $\underline{F}(4,45) = 2.73$, $\underline{p} < .05$. Only the feature and 250 album condition differed significantly by the Student Newman-Keuls multiple-comparison test, $\underline{p} < .05$. However, the difference in the number of photos examined between feature and album systems, on average 229 versus 75, was only marginally significant, a priori $\underline{t}(48) = 1.84$, $\underline{p} < .07$.

While we did not explicitly measure task completion time, album users spent more time per search. With two exceptions, feature users completed their seaches within 5-10 min. Most album users, in contrast, took one to three hours to finish.

3.2. Album Performance

Identification performance for album users deteriorated significantly as the number of photos examined increased (see Table 3). The hit rate decreased significantly from approximately 70% when targets were among the first 300 photos examined to 30% when 400 to 500 photos were examined, $\chi^2(1)$ for regression = 4.76, $\underline{p} < .05$. [The conventional χ^2 must be strengthened when the levels of the independent variable, in this case position of target photo, increase systematically. The appropriate test in this case is a modified χ^2 test for linear regression (Cochran 1954).]

The rate of incorrect identification (false alarms) also increased linearly with target position, $\chi^2(1)$ for regression = 4.64, $\underline{p} < .05$. Incorrect identifications averaged around 27% when the number of photos examined was less than 300, but increased to 70% for larger numbers of photos.

3.3. Feature Performance

On average, feature users examined a median of only 16.5 photos before encountering the photo of the target suspect. Mean retrieval rank (75.2) was higher because system performance was skewed. The only feature user to fail to identify the correct suspect also had the highest retrieval rank (472).

4. CONCLUSION

The purpose of this experiment was to compare the relative effectiveness of our computerized feature method for identifying suspects and the traditional mugshot album method. The feature method was superior. Subject witnesses identified the correct suspect in 90% of the feature searches but only 60% of album searches. The most dramatic difference in performance between the two methods, however, was in the rate of false alarms (0% for feature and 38% for album). Album users consistently identified the wrong person as the suspect. In contrast, feature users never did.

These comparisons overestimate the effectiveness of the album method. Presentation position of target suspects' photos were not randomly determined for the album method. The design of the present experiment dictated presentation of target suspects' photos in positions 1 through 300 and 400-500 only. If targets had been presented in the last 500 positions as well, album performance would have, at best, been no better than that of our subject witnesses presented targets in the range 400 to 500. Performance of those subjects had deteriorated to a 30% probability of correct identification and a 70% probability of identifying the wrong man as the offender. Extrapolation, with all its attendant problems, suggests identification performance of album users would average about 44% for hit rate (versus 90% for feature) and 55% for false alarms (versus 0% for feature).

Why these differences in performance between album and feature methods? We argue they are attributable to differences in the number of photos examined. The only difference between the feature and album methods in our experiment was in the sequencing of photos displayed to witnesses --

essentially random for album and rank ordered by similarity to witness descriptions of offenders for feature. Differences in sequence alter the position of target photos. Photos of target suspects were consistently among the first 50 photos examined in the database of 1000 photos (median retrieval rank = 16.5).

For the album system, identification performance deteriorated as the number of photos examined increased. The rate of correct identification of suspects decreased from approximately 70% to 30% while the rate of incorrect identifications increased from approximately 27% to 70%. Using the common χ^2 strengthened for testing linear relationships (Cochran 1954), we found a significant linear relationship between number of mugshots examined and the rate of false alarms and a significant negative relationship with hit rate. We reanalyzed the Ellis et al. (1989) data using the same statistical testing procedure. For typical faces, the results were the same as ours: the rate of false alarms increased with the number of photos examined ($\chi^2(1) = 4.67$, $\underline{p} < .05$) while the hit rate decreased systematically with the number of photos examined ($\chi^2(1) = 5.60$, $\underline{p} < .02$). uThese results replicate one of the most well established empirical effects in suspect identification (Deffenbacher, Carr, and Leu 1981; Laughery et al. 1971, 1974; Lenorovitz and Laughery 1984; Davies et al. 1979).

Our results extend those of Ellis et al. (1989) in several ways. First, they claim subjective feature systems won't work. Contrary to their assertion, subjective feature systems work at least as well as physical systems. Second, our database consists of official mugshots of offenders whereas theirs consists of photos of non-offenders. Third, we assess the similarity between witness and database rater descriptions of suspects using a Euclidean distance metric whereas Ellis et al. (1989) use a matching metric. [Not explicitly defined in their paper but based on a count of the percentage of features coded identically by both witness and database rater for each suspect in the database (Shepherd 1994 personal communication).] Both metrics work well though in several experimental tests we have found Minkowski distance metrics, and in particular Euclidean, to be superior (Lee et al. 1993). Fourth, since both our album and feature systems were computerized, differences in system performance can be confidently ascribed to the difference in sequencing of mugshots presented to witnesses. In contrast, Ellis' album system was not computerized whereas his feature system was. Consequently, the observed differences in performance could be attributed either to differences in mode of presentation of the mugshots (photographs versus digitized images displayed on screen) or to differences in mughsot sequencing.

Though they consistently outperform album systems, feature systems (subjective or physical) perform below optimum. Optimum performance in the present case would be a retrieval rank of one (i.e., a target suspect's photo would be the first mugshot from the database presented for examination to a witness). Only 1 of 10 feature system searches was this successful in the present experiment. We are currently testing ways of improving on this level of performance.

REFERENCES

Cochran, W.G. (1954) Some methods for strengthening the common χ^2 tests, *Biometrics*, *10*, 417-451.

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

Deffenbacher, K., Carr, T.H. and Leu, J.R. (1981) Memory for words, pictures and faces: Retroactive interference, forgetting and reminiscence, *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 299-305.

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

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

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

Laughery, K.R. and Wogalter, M.S. (1989) Forensic applications of facial memory research, in A.W. Young and H.D. Ellis (Eds.), *Handbook of research on face processing* (Elsevier, North Holland).

Lee, E.S., Densmore, H., and Whalen, T. (1993) Suspect identification by features, *Contemporary Ergonomics 1993 (The Ergonomics Society Annual Conference)*, (Taylor and Francis, London), 58-63.

Lee, E.S. and Whalen, T. (1993) 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) 494-499.

Lee, E.S. and Whalen, T. (1994a) Computerized feature retrieval of images: Suspect identification, *Ergonomics*, in press.

Lee, E.S. and Whalen, T. (1994b) Feature approaches to suspect identification: The effect of multiple raters on system performance, *Ergonomics*, in press.

Lenorovitz, D.R. and Laughery, K.R. (1984) A witness-computer interactive system for searching mug files, in G.Wells and E. Loftus, *Eyewitness testimony*. (Cambridge University Press, New York).

Whalen, T., Lee, E.S., and Baigent, G. (1994) A Computerized feature approach to suspect identification: Empirical tests with live suspects, *Proceedings of the 12th Triennial Congress of the International Ergonomics Association: Ergonomics and the Workplace*, Toronto, Canada, 15-19 August 1994, Vol. 5, 45-47.

Table 1

Example Features

Overall shape of face
1. Short vs long
2. Narrow vs broad
3. Bony vs fleshy
4. Not round vs round
5. Not well vs well proportioned
6. Weak vs strong facial structure

Table 2

Feature Versus Album System Performance (Success Rates for Identifying Suspects)

Performance	Feature	Album
Subject witnesses	n=10	n=40
Hits	90%	60%
False alarms	0%	37.5%
Misses	10%	2.5%

Table 3

Album Performance As a Function of Position in Album of Target Suspect's Photo

	Mean Position of Target in Album			
Performance	50	150	250	450
Hits	7	8	6	3
False alarms	3	2	3	7
Misses	0	0	1	0
Subject witnesses	n=10	n=10	n=10	n=10