38

AN EXPERIMENTAL STUDY OF A GESTALT BASED DATABASE FOR MUG SHOTS

Unni Astad, Frank Safayeni **, and Darrell Raymond ****

* Norwegian Computing Center
P.O. Box 114 Blindern, N-0314 Oslo, Norway.
** Department of Management Sciences, *** Department of Computer Science
University of Waterloo, Ontario, N2L 3G1 Canada

*E-mail: unni.astad@nr.no

KEY WORDS: Suspect identification, mug shots, image retrieval, database.

ABSTRACT: The paper describes a series of experiments testing a recognition based computerized mug shot database system (gestalt based) for police use. The prototype is compared to the more traditional feature based database system. The gestalt based database was shown to perform significantly better than the feature based database when the quality of the signals was improved. In addition, the gestalt based system outperformed the feature based system in terms of efficiency of each question and ease of use. Issues concerning the coding of the database and the use of a hybrid system are also discussed.

1 INTRODUCTION

At many suspect identification interviews, a witness, who has briefly seen a suspect, is asked to identify the person by looking through a selection of mug shots (photographs of suspects) from a police mug shot file (album method). Both published work and personal conversation with the police suggests that this system has not proven particularly successful (Ellis, Shepherd, Shepherd, Klin, & Davies, 1989). The main problem is that witnesses can be easily confused, and if the suspect does not appear among the first 100 to 200 mug shots viewed, the witnesses have problems making a correct identification (Davies, Shepherd, & Ellis, 1979; Laughery, Alexander, & Lane, 1971; Laughery, Fessler, Lenorovitz, & Yoblick, 1974).

Mug shots can be stored in computer databases. This is of great value if a computerized system for the storage and retrieval of mug shots can outperform the album method in terms of the number of mug shots a witness has to view before the correct mug shot (the suspect) appears. The problem is to devise a system that can efficiently retrieve mug shots based on the memory of a witness. Most witnesses have only a visual memory of a person's face. A retrieval system must therefore be able to map the amount and type of information available in the witness' memory to the information in the mug shot database. The major system constraint is therefore on the retrieval side.

2 FEATURE BASED APPROACH

The few research groups that have worked with the design of such database systems have been pursuing the feature based approach (Harmon, 1973; Ellis et al., 1989; Lee & Whalen, 1993; Lee, Whalen, Samoluk, & Densmore, 1994). In a feature based model, single features of the face on a mug shot are rated, either by a rater or by objective measurements. For example, attributes such as the blueness of the eyes or the length of hair are evaluated. During retrieval, the witness makes these types of ratings, which are then compared to the ratings in the database. Faces that have similar

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

ratings are then retrieved in order of similarity, and are shown to the witness.

A feature based system requires the witness to recall single features. The problem with this approach is that its success rests on the assumption that a witness has a detailed memory of facial features. There is little support in the literature for this assumption. Evidence from research on the processing and memory of facial information suggests that people are generally better at recognizing whole faces than recalling single features (Hunter, 1973; Neisser, 1976; Walker-Smith, 1978; Laughery, Duval, & Wogalter, 1986; Deffenbacher, 1989).

Another important aspect is the forensic situation itself, which impairs the processing of facial information. Factors such as the insignificance of the event, less than ideal observation conditions, a short period of observation, witness stress, and so on, may all reduce the encoding of facial details in the witness' memory (Buckhout, 1982). This implies that the existing feature based database systems may not match properly with the facial memory inherent in people. A critique of reported experiments is that the success of these systems are due to the subjects knowing beforehand that they would be asked to recall single features. Subjects therefore paid attention to features they would not have noticed in a real situation.

3 GESTALT BASED APPROACH

People are generally good at recognizing faces, but not good at recalling single features or minor details. The alternative approach is therefore recognition based. In such an approach, both the storage and the retrieval of mug shots are based on perceived similarity to the whole face rather than to single features. The database is called a gestalt based database. The term *gestalt* refers to treating the face as an integrated whole, more than the sum of its parts.

Every mug shot in a gestalt based database is coded according to its similarity to a predefined set of faces. The witness is asked to judge how much the faces resemble the image of the suspect in their memory. These ratings are then used to extract similar faces from the database. The system draws on the individual's strength of recognition rather than recall.

4 EXPERIMENTAL STUDY

A prototype gestalt based database was evaluated and refined in a series of experiments. The purpose of the experiments was to compare the gestalt based database to an existing feature based database, and to improve the performance of the gestalt based database. The study consisted of two main parts. In the first part both the gestalt based database and a feature based database were tested. In the second part the effect of various matching strategies, such as noise removal and training was tested on the gestalt based database. The experimental condition, which was the same throughout the experiments, was set up to ensure a more realistic situation. The success of the systems was measured in terms of how many mug shots the subjects (witnesses) had to look at before the target was found. The measure is called retrieval rank, and is given in both actual numbers and percentage of the whole database. The lower the number or percentage, the better the retrieval system.

4.1 Prototype Gestalt Based Database

The database contained 298 official mug shots of male Caucasians aged 18 to 33 years. The 298 mug shots were randomly sampled from an original database of 1000 mug shots. The facial similarity in the database was high since variables such as race, sex, and age, were eliminated. The mug shots were colour photos taken under standard condition: a frontal view of the face taken from the shoulder up (90x125 mm prints). The photos were digitized. The final image size as it appeared on the computer screen was 45x53 mm, and the colour resolution was 8 bits per pixel.

Each picture in the database was given a ten element code, where each element represented a similarity rating. This rating was the outcome of a similarity question posed as a choice between two faces. The ten pairs of faces, which constituted the ten questions, were sampled in the following manner: A number of faces were randomly put together to form pairs. From these pairs, a subset of pairs were selected in which the two pictures in a

pair were as different from each other as possible. This was done to make the similarity task easier, because two very dissimilar faces were assumed to be easier to distinguish than two similar faces.

Four raters rated all the 298 pictures in the database with respect to perceived similarity to the ten pairs of faces. Both the pairs and the pictures being rated were presented to the raters on the computer screen. The judgments from all four raters were recorded and added up so that each picture in the database had a code consisting of 10 elements corresponding to the similarity rating on each of the 10 pairs. A typical storage code looked like: "0.75R 1.0L 0.75L 0.5RL 0.5RL 1.0R 1.0R 1.0L 0.5RL 0.5RL". The number in front of the letter indicated the percentage of raters who agreed on the decision, while the letter said whether the choice was the left hand picture (L) or the right hand picture (R). To reduce the complexity, the 1.0 values were combined with the 0.75 values to form one signal (coded L or R), while the 0.5 values were eliminated (coded X). Using the typical code presented above, the final code which the witness had to match looked like: "R L L X X R R L X X". The witness would do comparable similarity judgments based on their memory of the target picture. These ratings were then matched to the database, and mug shots were selected in order of number of matches on the R and L values.

4.2 Feature Based Database

The feature based database used was Lee's database (Lee et al. 1993, 1994). The 298 mug shots in this database were the same as in the gestalt based database, but with a feature coding. All the pictures were coded by one rater on 107 features using a 5-point scale. The search in the database was done by matching the witness' coding and the raters' coding.

4.3 Methodology

All the subjects were naive subjects; that is, they had no knowledge of the research topic.

The basic stimuli was a story in which the target picture appeared among four pictures of buildings. The subjects were told that after the story was finished they would be asked to answer some questions related to the story, but they were not told which part of the story. This was to ensure that the facial information encoding was natural and automatic, rather than deliberately to recall the person. The set-up was similar to a forensic situation in which the witness is not aware of a crime being committed, but at a later stage has to identify the criminal. All the subjects were told the same story and saw the same pictures, except for the target picture which varied across the subjects. A total of twelve targets were randomly chosen from the database of 298 pictures.

The data collection instrument varied depending on whether the feature based database or the gestalt based database were being accessed. To access the feature based database, the subjects had to fill out a questionnaire about the features of the target. The subjects were told to skip those questions they did not know. The remaining answers were used to access Lee's feature based database. To access the gestalt based database, the subjects were shown ten pairs of faces on a computer screen, one at a time. For each pair the subjects had to judge whether they perceived the person in the story to be more similar to the picture on the left or to the picture on the right. When they had made their decision, the next pair was shown on the computer screen. The subjects' ten binary decisions were then used to access the gestalt based database.

4.4 Results of the Initial Experiment

The purpose of the first part was to compare the feature based technique for storing and retrieving mug shots to the gestalt based technique. Twenty-four subjects participated in this part. All subjects accessed both databases. The subjects' responses on the questionnaire were used to access the feature based database, and the subjects' similarity judgments using the computer were used to access the gestalt based database. In both cases the retrieval rank of the targets were reported. The result is shown in Table 1.

N = 24	Feature	Gestalt
$\overline{X}$	57 (19%)	119 (40%)
S_x	51	88

Table 1: Initial version of a gestalt based database and a feature based database.

The feature based system had a mean retrieval rank that was significantly lower than the gestalt based database (Wilcoxon signed rank test, 1-tailed p = 0.0043). The gestalt based database did not perform significantly better than randomly assigned codes, having an average retrieval rank of 40%.

As expected, the feature based database performed poorly when the experimental condition was more like a recognition condition than a recall condition. Lee et al. (1993, 1994) report average retrieval rank of target suspect using one rater, as 1.6% and 5.6% of a database. This increased to 19% in this study. But even so, the result of the experiment was surprising. One had expected that the gestalt based database would perform better than the feature based database under the present experimental condition. However, when the subjects evaluated the two tasks with respect to ease of use, 83% of the subjects preferred the gestalt based system.

4.5 Improving the Gestalt Based Database

The purpose of the second part of the study was to improve the performance of the initial gestalt based database by testing the effect of various matching strategies, like noise removal and training. Thirty-two subjects participated in this part.

There are two main sources of noise in a database system. One source of noise is the *witness*, and another is the *database* itself.

There are two ways of removing noise from the witness. One is to eliminate those decisions the witness feels uncertain about. Another is to train the witness in how the raters did their similarity judgments, thereby increasing the likelihood of a better mapping between the raters codes and the witness' judgments. Reducing the noise from the gestalt based database by asking people to eliminate their unsure decisions did not result in a reduced retrieval rank. By comparison, the feature based database is quite successful in removing witness noise, by simply instructing witnesses not to answer the questions they do not remember. Whole faces seem to differ from single features in this respect. The reason for this may be the complexity of whole faces as compared to single features. Similarly, training the subjects in how the coding in the database was done did not help them to make better similarity judgments. This result was surprising. It was initially believed that training the subjects would increase the number of matches. The subjects typically commented that the training confused them, and that they did not know whether they should base their decisions on their own judgments or that of the raters.

There are many methods which have potential to improve the codes. One is to ask better questions, that is, to use pairs of pictures which better represent the content of the database, which are easier to make a similarity decision on, and so on. However, this method of reducing the noise in the database was not considered at this stage, because it meant recoding and retesting the whole database. But there was another way of reducing the noise without altering the initial coding. That was to purify the codes using only those signals that all four raters agreed on (1.0R &1.0L), and ignore the rest; that is, eliminating the codes with values of 0.75 and 0.5. The subjects were more likely to agree with the raters when all four raters agreed (1.0), than when only 3/4 of the raters agreed (0.75). The average percentage of R & L matches between the subjects and the raters was 56% for the combined 1.0 and 0.75 codes, and 73% for the 1.0 only codes. So by counting matches only on the strong signals, the probability of error decreased. The average number of signals in the codes for the twelve targets was reduced to 4.1. Purifying the codes in this way did not in the end improve the performance of the gestalt based database. There was still too much noise in the database.

The codes could be further improved by using more raters to code the pictures in the database, thereby reducing the probability of a wrong signal. But since this involved additional coding of all the pictures, a simplified version was chosen. Making use of the rating from all the subjects as a form of feedback in the system, the coding on the twelve targets was improved. After including the subjects similarity judgments as ratings, the average number of raters on each of the twelve targets was 12.4 raters. Using a significance level of $\alpha = 0.10$ for the signals, new codes for the twelve targets were created. For an element in the code to be viewed as a strong signal, less than 10% of the raters and the subjects could disagree with the choice. The

percentage of matches between the subjects and these signals was 88%, and the average number of signals in the codes for the twelve targets was 5.2. So by purifying the codes in this way both a better mapping between the subjects and the codes and more signals were achieved. Table 2 shows the average retrieval rank for the combined fifty-six subjects using the gestalt based database with improved signals, $\alpha = 0.10$. The first column shows the result using all twelve targets, while the second column shows the result when one outlier target was removed. The outlier target had only one signal in its code, resulting in an extremely high retrieval rank.

	Gestalt	
	N = 56	N = 51
$\overline{X}$	55 (18%)	35 (12%)
S_x	75	38

Table 2: Gestalt based database using improved signals ($\alpha = 0.10$)

These results were compared to the result of the feature based database as presented in Table 1, using the Wilcoxon-Mann-Whitney Test. The following were found: Including the outlier, there was no significant difference between the two techniques in terms of retrieval rank. However, when this picture was removed, the average retrieval rank using the gestalt based database was significantly lower than the average retrieval rank using the feature based database (1-tailed p = 0.0233). The improved gestalt based database performed very well for 92% of the pictures having an average retrieval rank of 12% of the database.

Comparing the two methods in terms of efficiency of the questions being asked, the feature based database used 107 questions, while the gestalt based database used only ten to achieve the same average retrieval rank. Therefore the questions in the gestalt based database were ten times as efficient as those in the feature based database.

5 DISCUSSION

Based on the major findings in this research, there are two classes of issues worth discussing further: The coding of the gestalt based database and a hybrid system using both gestalt and feature properties.

5.1 Coding of the Gestalt Based Database

The goal of improving the signals was to achieve a better mapping between the subjects and the codes in the database. Good performance resulted when the subjects' own coding were used to improve the signals. A minor pilot experiment indicated that similarity judgments based on perfect information (raters) generated slightly different signals than similarity judgments based on memory (witnesses). If the lack of success using the raters' signals was due to these signals being based on perfect information which do not map very well onto the subjects memory, then this has implications for how the coding in the database should be carried out. By generating codes based on perfect information one assumes that the witness' memory of the same information is similar. The mistake in such an assumption is analogous to the wrong assumption underlying the feature based database, that by storing the mug shots based on features one assumes that the memory of the face is in terms of single features. In both cases, it is the mapping between the codes and the witness' memory that is not good enough. Therefore, rating the prototype database by memory to achieve a better mapping between the codes in the database and the witness' memory should be tested.

A second issue is the quality of the ten questions/pairs. During the experiments it was noted that some pairs of faces work better as questions than others. It would therefore be useful to select pairs and pretest these with respect to ease of getting pure signals, how well they divide the database, independence of other pairs, and so on.

A final issue is the effect of using more questions than the ten used in this prototype. Issues of concern here is the signal/noise ratio. Adding more questions increases the possibility of both more signals and more noise.

5.2 Hybrid System

It would be interesting to pursue the design of a hybrid system, in which both features and overall similarity judgments are used to retrieve pictures from the database. In the course of collecting the

data for the experiments, evidence was found that subjects notice some overall or gross features. All the subjects could recall hairstyle and facial hair. The album method currently in use by the police is such a hybrid system, in which a witness is first asked for a number of gross features. When no more features can be recalled, the witness is presented with a number of mug shots.

By using the two gross features, hairstyle and facial hair, in addition to the gestalt based database, the subjects needed only to look at 8% of the database. The effect of adding these two features was greatest for those pictures that did not perform well in the gestalt based database. By using such a hybrid system, faces which are not easily remembered in terms of features may be better captured by similarity judgments and vice versa.

REFERENCES

Buckhout, R. (1982). Eyewitness Testimony. In U. Neisser (Ed.), *Memory Observed*, W. H. Freeman and Company, USA

Davies, G., Shepherd, J.W., & Ellis, H.D. (1979). Similarity Effects in Face Recognition. *American Journal of Psychology*, 92, 507-523.

Deffenbacher, K.A. (1989). Forensic Facial Memory: Time is of Essence. In: A.W.Young & H.D. Ellis (Eds.), *Handbook of Research on Face Processing*, Elsevier Science Publishers B.V. (North-Holland).

Ellis, H.D., Shepherd, J.W., Shepherd, J., Klin, R.H., & Davies, G.M. (1989). Identification From a Computer-Driven Retrieval System Compared with a Traditional Mug-Shot Album Search: A New Tool for Police Investigations. *Ergonomics*, 32, 167-177.

Harmon, L.D. (1973). The Recognition of Faces. *Scientific American*, 229, 71-78.

Hunter, I.M.L. (1964). *Memory*. Penguin Books. U.K.

Laughery, K.R., Alexander, J.F., & Lane, A.B. (1971). Recognition of Human Faces: Effects of Target Exposure Time, Target Position, Pose Position, and Type of Photograph. *Journal of Applied Psychology*, 55, 477-483.

Laughery, K.R., Duval, C., & Wogalter, M.S. (1986). Dynamics of Facial Recall. In: H.D. Ellis, M.A. Jeeves, F. Newcombe, & A. Young (Eds.), *Aspects of Face Processing. NATO ASI Series*, Martinus Nijhoff Publishers, Dordrecht.

Laughery, K.R., Fessler, P.K., Lenorovitz, D.R., & Yoblick, D.A. (1974). Time Delay and Similarity Effects in Facial Recognition. *Journal of Applied Psychology*, 59, 490-496.

Lee, E.S. & Whalen, T. (1993). Computer Image Retrieval by Features: Suspect Identification. *Proceedings of the INTERCHI'93.*

Lee, E.S., Whalen, T., Samoluk, S., & Densmore, H. (1994). Empirical Tests of a Computer Suspect Identification System: An Alternative to Mugshot Albums. *Second Biennial European Joint Conference on Engineering Systems Design and Analysis*, University of London, London, U.K., 4-7 July 1994.

Neisser, U. (1976). *Cognition and Reality*, W. H. Freeman and Company, U.S.A.

Walker - Smith, G.J. (1978). The Effects of Delay and Exposure Duration in a Face Recognition Task. *Perception and Psychophysics*, 24, 63-70.