FINAL REPORT

EVALUATING A NEW TECHNIQUE FOR IMPROVING EYEWITNESS IDENTIFICATION.

Victor S. Johnston
Department of Psychology
New Mexico State University.
Las Cruces, New Mexico 88003

NIJ Grant #
90-IJ-CX-0025

FINAL REPORT

EVALUATING A NEW TECHNIQUE FOR IMPROVING EYEWITNESS IDENTIFICATION.

Victor S. Johnston
Department of Psychology
New Mexico State University.
Las Cruces, New Mexico 88003

SUMMARY REPORT.
EVALUATING A NEW TECHNIQUE FOR IMPROVING EYEWITNESS IDENTIFICATION

Victor S. Johnston.
Department of Psychology.
New Mexico State University.
Las Cruces, New Mexico 88003

The primary goal of this study was to evaluate a facial recognition system in which facial composites were constructed by witnesses using a computer. The computer system contained facial features that permitted over 34 billion composites to be generated. The system initially generated 20 possible faces that were rated by a witness for their general resemblance to a culprit. These ratings were then used to guide a genetic algorithm (GA) search process which generated a new set of 20 faces (second generation) based upon the most highly rated composites of the first generation. The new faces were again rated for resemblance to the culprit. This process continued until the culprit's face was evolved.

A second goal was to determine the best settings for variables which influence the efficiency of the GA search. These variables included: (a) the genetic coding system (e.g binary or Gray code), (b) the GA parameters (e.g. mutation and crossover rate), (c) the number of generations needed, and (d) the user interface.

Of the different coding systems, binary was found to have the best characteristics for rapid evolution. The optimal parameters (mutation and crossover rate) were evolved using a meta-level simulation program. Simulations also demonstrated that a close resemblance to a culprit's face could be evolved in ten generations when accurate fitness feedback was provided and the user interface provided a means for "freezing" highly desirable facial features.

The GA process was evaluated experimentally by requiring subjects to evolve the facial composite of different culprits at varying delays after exposure to a simulated crime. The quality of these composites were evaluated by the ability of independent judges to correctly identify the culprit from an array of faces, using only the evolved composites. These studies revealed that a subject's ability to recognize a culprit varied with the facial characteristics of the culprit. Prototypical faces were poorly recognized, but faces with distinctive features were well remembered up to one week after exposure. When recognition was good, the GA procedure evolved composites which were identified by judges on more than 50% of cases.

TABLE OF CONTENTS:

TOPIC. PAGE

A. INTRODUCTION. 1

B. GOALS. 1

C. REVIEW OF THE LITERATURE. 2
(1) Current Identification Procedures. 2
(2) Cognitive Processes. 3
(3) The GA Strategy. 7

D. WORK COMPLETED. 7

E. RESEARCH PLAN. 10
(1) Design Phase. 10
(2) Testing Phase. 11

F. RESULTS 11
(1) Design Phase 11
(a) Stochastic Universal Sampling. 12
(b) Gray Code and Binary Code. 13
(c) Optimization of GA Parameters. 15
(d) Evaluating the User Interface. 15
(e) Program Modifications. 20
(f) Evaluating Freeze and Flood Options. 21

(2) Testing Phase. 23
(a) Culprit Selection. 23
(b) Pilot Study 2 25
(c) Pilot Study 3 27
(d) Evaluation of the FacePrints Process. 27

G. DISCUSSION. 35

H. CONCLUSIONS. 36

I. EXPECTED BENEFITS 37

J. REFERENCES. 39

EVALUATING A NEW TECHNIQUE FOR IMPROVING EYEWITNESS IDENTIFICATION.

Victor S. Johnston
Department of Psychology
New Mexico State University.
Las Cruces, New Mexico 88003

(A) INTRODUCTION:

Often the single most important piece of evidence available to law enforcement officers is the description of a suspect by an eyewitness. Although humans have excellent facial recognition ability, they often have great difficulty recalling facial characteristics in sufficient detail to generate an accurate composite of the suspect. As a consequence, current identification procedures, which depend heavily on recall, are not always adequate.

Unlike current procedures, a genetic algorithm (GA) is capable of efficiently searching a large sample space of alternative faces and of finding a "satisficing" solution in a relatively short period of time. Since such a GA procedure can be based on recognition rather than recall, and makes no assumptions concerning the attributes of witnesses or the cognitive strategy they employ, it should be able to find an adequate solution irrespective of these variables. This report is an initial evaluation of such a GA based computer program: FacePrints.

(B) GOALS:

The primary goal of this study is to explore the use of a genetic algorithm as an alternative method for evolving an accurate facial composite. This goal will be evaluated experimentally by requiring subjects to evolve the facial composite of different culprits at varying delays after exposure to a simulated crime. The quality of these composites will be evaluated by the ability of independent judges to correctly identify the culprit from an array of faces, using only the evolved composites.

A second goal of the current study is make an initial determination of the best settings for variables which influence the efficiency of the GA search. These variable include: (a) the genetic coding system (e.g binary or Gray code), (b) the GA parameters (e.g. mutation and crossover rate), (c) the number of generations used, and (d) the user interface. These settings will then be used to achieve the primary goal.

(C) REVIEW OF THE LITERATURE:

(1) Current Identification Procedures:

The human face can convey an incredible quantity of information. As Davidoff (1986) has noted, age, sex, race, intention, mood and well-being may be determined from the perception of a face. Additionally, humans can recognize and discriminate between an "infinity" of faces seen over a lifetime, while recognizing large numbers of unfamiliar faces after only a short exposure (Ellis, Davies and Shepherd, 1986).

When the nature of the perceiver is fixed, such as when a witness is required to identify a criminal suspect, only the configuration and presentation of the stimulus face may be varied to facilitate recognition. To ensure success under these circumstances, the facial stimuli must provide adequate information, without including unnecessary details that can interfere with accurate identification. A body of research has attempted to uncover the important factors governing facial stimuli and methods of presentation that are most compatible with the recognition process.

The most systematic studies of facial recognition have been conducted in the field of criminology. Beyond the use of sketch artists, more empirical approaches have been developed to aid in suspect identification. The first practical aid was developed by Penry (1974), in Britain, between 1968 and 1974. Termed 'PhotoFit, this technique uses over 600 interchangeable photographs of facial parts, picturing five basic features: forehead and hair, eyes and eyebrows, mouth and lips, nose, and chin and cheeks. With additional accessories, such as beards and eyeglasses, combinations can produce approximately fifteen billion different faces. Initially, a kit was developed for full-face views of Caucasian males. Other kits for Afro-Asian males, Caucasian females and for Caucasian male profiles soon followed (Kitson, Darnbrough and Shields, 1978). Alternatives to PhotoFit have since been developed. They include the Multiple Image-Maker and Identification Compositor (MIMIC), which uses film strip projections; Identikit, which uses plastic overlays of drawn features to produce a composite resembling a sketch, and Compusketch, a computerized version of the Identikit process. The Compusketch software is capable of generating over 85,000 types of eyes alone (Visitex Corporation, 1988). With no artistic ability, a trained operator can assemble a likeness in 45 to 60 minutes. Because of such multiple advantages, computer aided sketching is becoming the method of choice for law enforcement agencies. As of May 1st, 1988, fifty Compusketch programs were in use in eighteen states in the USA.

Because of its wide distribution, the PhotoFit system has generated the largest body of research on recognition of composite facial images. Ellis, Davies and Shepherd (1978) have compared memory for photographs of real faces with memory for PhotoFit faces which have noticeable lines around the five component feature groups. They have reported that subjects recognize the unlined photographs more easily. The presence of lines appears to impair memory, and random lines have the same effect as the systematic PhotoFit lines. In a second paper, they also note that individuals display a high degree of recognition of photographs, but describe a human face poorly (Davies, Shepherd and Ellis, 1978). They contend that at least three sources of distortion arise between viewing a suspect and a PhotoFit construction: 'selective encoding of features', 'assignment to physiognomic type', and 'subjective overlay due to context'. These distortions contribute to the production of caricatures of a suspect rather than accurate representations. Hagen and Perkins (1983) have compared true line-drawn caricatures with profile-view and three-quarter-view photographs. Their research indicates that facial recognition is good within a medium, but is seriously disrupted when changing to another medium, especially those involving caricatures. These results suggest that unadulterated photographs are superior to caricatures for the identification of real faces.

A detailed study by Laughery, Fowler and Rhodes (1976) has shown that beards and hair styles contribute significantly to errors, but the presence or absence of glasses has little effect on recognition. Additionally, they note that helping an artist sketch a picture of a suspect, or using an identification kit, can assist recognition even after periods of six months to one year. The effects of delay on recognition and PhotoFit construction of faces has also been examined by Davies, Ellis and Shepherd (1978). After three weeks of elapsed time, they found no detectable decrease in reconstruction accuracy. Hall (1976) has demonstrated that memory for faces is very good if no external sources of bias are introduced. However, biasing instructions, intensive rehearsal of the suspect's appearance with verbal feedback, or too intense concentration on minor facial details can impair performance. Loftus and Greene (1980) have also demonstrated this susceptibility to interference by successfully misleading subjects with questions or oral descriptions of the face from another source. Ideally, to avoid such bias, an unskilled witness should be able to generate a composite facial stimulus unaided and uninfluenced.

(2) Cognitive Processes:

The concept of a schema is central to almost all theories concerned with the encoding, storage or retrieval of facial information. This idea can be traced back to Bartlett (1932), who defined a schema as "an active organization of past reactions, or of past experience..." (p. 201). For Bartlett, information processing occurred when new information interacted with old information, contained in a schema, and this integration accounted for memory distortion (Brewer and Nakamura, 1984). It is a consequence of Bartlett's "reconstructive" model that no episodic representation of an original event is left intact, causing memory for faces to become more distorted over time as they come to resemble a stereotypical face.

Posner's (1973) prototype theory is also a schema model, where the average mean value of feature information is included in a facial prototype. In this "averaging model", the prototype features depend on the central tendencies of the distributions of features experienced. The most frequently seen features may not necessarily be the features contained in the prototype. Neumann's (1974) alternative "attribute frequency model" of prototype formation hypothesizes that the modal features will be extracted and included in the prototype. To find whether the "averaging model" or the "attribute frequency model" provides the best description for the data, Neumann (1977) used a bimodal distribution of features as stimuli. If averaging leads to prototype formation then the central tendency of the features should promote recognition accuracy; if attribute frequency is the important prototype determining factor, then faces possessing features from the extremes of the distribution should be recognized better. In fact, Neumann found support for both models, the results depending on whether or not subjects had information on which features varied in the study set.

Solso and McCarthy (1981) have provided evidence supporting prototype formation from frequently shown features by demonstrating the predicted effects of such a prototype on recognition judgements. After exposure to a variety of Identikit faces, subjects were tested for recognition using the original faces, new faces, and a prototype composed of the most frequently seen features. With high confidence, subjects consistently misidentified the prototype face as previously having been seen. These results persisted over many weeks and provide strong evidence for the use of an hypothesized facial prototype. Light, Kayra-Stewart, and Hollander (1979) have supported the influence of a general facial prototype by showing that faces rated as "typical" were recognized less accurately than faces rated as "unusual". They argue that the more similar a target face is to the prototype, the less likely it is that the target face will be accurately recognized. Valentine and Bruce (1986) have extended this research by comparing the recognition of distinctive familiar faces with typical familiar faces. They found that distinctive familiar faces were more easily recognized than typical familiar faces and concluded that " the effects of distinctiveness arise because faces are encoded by comparison to a single prototypical face." (p 525). Haig (1986a,1986b), has presented evidence supporting the use of a stored prototype during facial recognition. This research suggests that memorization of a particular face occurs by placing unusual features or combinations of features onto a "bland prototypical face". Subjects used the head outline, followed by the eye and eyebrow grouping, and then the mouth and nose, during the recognition process. Based on this research Haig concluded that facial recognition is a two stage process. The first results from the use of a bland, almost featureless prototypical face structure while the second involves the mapping of specific features onto this prototype.

Other studies of feature saliency appear to support Haig's findings. Cook (1978) investigated the eye movements and fixations of subjects while examining both familiar and unfamiliar faces. Following an initial encoding exposure, three or four fixations were found to be necessary for the recognition of unfamiliar faces; the modal time was 0.9 secs. With no prior exposure, familiar faces (e.g. Paul Newman or Gerald Ford) required approximately four fixations; the modal time was 1.3 secs. For both conditions subjects used the eyes, nose, and mouth, respectively, in order to achieve recognition. Cook concluded that recognition is achieved by examining significant features and comparing those features to a stored prototypical face.

In a study using a computer-implemented caricature generator, Rhodes, Brennan and Carey (1985) have found that caricatures of familiar faces are identified more quickly than either veridical•line drawings or anticaricatures (made from minimizing the distances which are exaggerated in caricatures). This again suggests that the distinctive aspects of a face may be represented in comparison with a generic norm.

Although there is substantial evidence for facial prototypes, connectionist models offer an alternative viewpoint on prototype formation and its use during the recognition process. Based on theoretical arguments and empirical findings, McClelland and Rumelhart (1985) have concluded that prototypes are not always used to categorize or influence judgements. From the connectionist viewpoint, specific exemplars overlap to form a prototype, and, "when all the distortions are close to the prototype, or when there is a very large number of different distortions, the central tendency will produce the strongest response, but when the distortions are few, and farther from the prototype, the training exemplars themselves produce the strongest activation " (p. 182). Thus, McClelland and Rumelhart argue for circumstances when either the prototype or the stored exemplars can serve as the basis for recognition judgements.

The complexity of the facial recognition process is further magnified by the involvement of gender and/or cerebral dominance factors. Going and Read (1974) have found that women subjects recognize both highly unique and non-unique female faces more frequently than male faces, while men recognize faces of both types and genders with equal facility. In addition, unique faces of both genders are correctly identified more often, with exceptional female faces recognized more frequently than unique male faces. Freeman and Ellis (1984) have related such gender based differences in performance to the cerebral asymmetry occurring during sexual development. They have found that males exhibit a left visual field (right hemisphere) superiority for rapidly viewed faces with low-detail. In fact, subjects of both sexes make fewer recognition errors and appear to extract more information from the left half of a face, even when both halves are mirror images (Kennedy, et al.,1985). Many studies have supported the conclusion that faces are more efficiently processed in the right cerebral hemisphere. However, Rhodes (1985) has found that the half-face of a famous person, when presented in the left visual field, is not retained in memory. He believes that this may be due to asymmetrical scanning or attentional factors beyond laterality effects.

Ross-Kossak and Turkewitz (1986) suggest that the direction of an individual's hemispheric advantage affects the type of information processing strategy used. They have found that omission of selected facial features degrades the performance of subjects with a left-hemisphere advantage, while inversion of faces impairs performance of subjects with a right-hemisphere advantage. Miller and Barg (1983) presented drawings of faces (assembled from Identikit transparencies) to a subject's right or left visual field. This technique revealed that discriminations along feature dimensions are more accurate for faces shown to the right visual field, while discriminations involving spatial relations among the eyes, nose and mouth are more accurate for faces perceived in the left visual field. Significantly, these results do not extend to line drawings of houses. In addition, processing strategy may change with increasing familiarization. Ross-Kossak and Turkewitz (1986) have found that subjects who begin with a left-hemisphere advantage shift to a right-hemisphere advantage across trials, while those beginning with a right-hemisphere advantage decrease and then increase the magnitude of this advantage over trials. Thus, the methods witnesses employ during face recognition reveal processing dichotomies between the cerebral hemispheres that appear to vary between analytical, feature based, and holistic, organizational strategies.

A major conclusion from the cognitive research is that the mechanics of Compusketch and its predecessors, PhotoFit, MIMIC, and Identikit, may actually inhibit recognition, by forcing witnesses to employ a specific cognitive strategy; namely, constructing faces from isolated feature recall. Since facial recognition appears to also involve holistic processes, the single feature methodology may be inappropriate. Indeed, Davies and Christie (1982) have suggested that the single feature approach may be a more serious source of recognition distortion than interference from an outside source.

Many of the problems and limitations of the existing identification systems may be eliminated by adopting a strategy for generating faces that exploits the well developed skill for facial recognition, rather than individual feature recall. Moreover, the approach may be designed so that it accommodates a wide variety of individual styles of cognitive processing. The proposed method is to use the genetic algorithm (GA) to generate composite faces, evolving a suspect's face over generations, and using recognition as the single criterion for directing the evolutionary process.

(3) The GA Strategy:

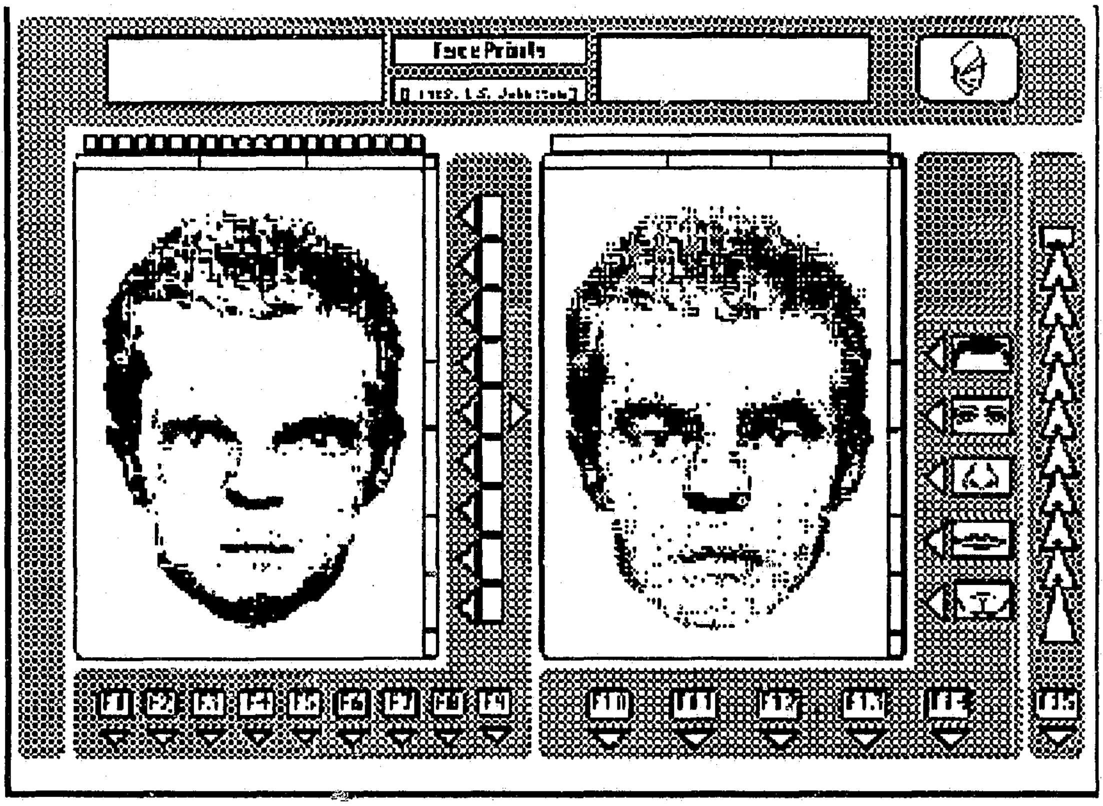
The simple GA, first described by Holland (1978), is a robust search algorithm based upon the principles of biological evolution. In essence, the GA is a simulation of the evolutionary process, and makes use of the powerful operators of "natural" selection, mutation and crossover to evolve a solution to any complex design problem. It is capable of searching a very large sample space and finding a "satisficing" (often optimal) solution in a relatively small number of generations. The following section describes the first attempt to use a GA in order to evolve a composite face. We call this facial composite process "FacePrints".

(D) WORK COMPLETED:

In the current design of FacePrints, a series of twenty faces (phenotypes) are generated from a random series of binary number strings (genotypes) according to a standard developmental program. During decoding, the first seven bits of the genotype specify the type and position of one of 32 foreheads, the next seven bits specify the eyes and their separation, and the remaining three sets of seven designate the the shape and position of nose, mouth and chin, respectively. Combinations of these parts and positions allow over 34 billion composite faces to be generated.

The position of all features are referenced to a standard eye-line, defined as an imaginary horizontal line through center of the pupils. Four positions (2 bits) are used to specify the vertical location of each feature (eyes, mouth, nose and chin) with reference to this standard, and two bits to specify pupil separation. These positions cover the range of variability found in a sample population.

The initial twenty random faces can be viewed as single points spread throughout the 34 billion point multi-dimensional "face-space". The function of the GA is to search this hyperspace and find the best possible composite in the shortest possible time. The first step in the GA is the “selection of the fittest” from the first generation of faces. This is achieved by having the witness view all twenty faces, one at a time, and rate each face on a nine point scale, according to any resemblance, whatsoever, to the culprit (a high rating signifies a good resemblance to the culprit).



Rating Scale for Composite   Best Face from Previous Generation

Figure 1. Example of composite generated by FacePrints (v3.0).

Figure 1 shows the display presented to the witness and the rating scale for entering the measure of relative fitness. This measure does not depend upon the identification of any specific features shared by the culprit and the composite face; the witness need not be aware of why any perceived resemblance exists. After ratings of fitness are made by the witness, a selection operator assigns genotype strings for breeding the next generation, in proportion to these measures. Selection according to phenotypic fitness is achieved using a "spinning wheel" algorithm. The fitness ratings of all twenty faces are added together (TotalFit) and this number specifies the circumference of an imaginary spinning wheel, with each face spanning a segment of the circumference, proportional to its fitness. Spinning a random pointer (i.e. selecting a random number between one and TotalFit) identifies a location on the wheel corresponding to one of the twenty faces. Thus, twenty random spins of the pointer select twenty breeders in proportion to their fitness.

Sexual reproduction between random pairs of these selected breeders is the next step in the GA. Since each breeding pair produces two offspring, the population size remains constant. Random breeding between the selected breeders employs two additional operators:

Crossover and Mutation. When any two genotypes mate, they exchange portions of their bit strings according to a user specified crossover rate (Xrate = number of crosses per 1000), and mutate (1 to 0, or 0 to 1) according to a user specified mutation rate (Mrate = number of mutations per 1000). Two selected genotypes (A and B) can be represented as shown below.

A 1,0,0,0,1,1,0,0,1,0,1,0,1,1,1,..........1
B 0,1,1,0,0,0,1,0,1,1,1,0,0,0,0,..........0

During breeding the first bit of A is read into newA and the first bit of B into newB. At this point a check is made, to see if a crossover should occur. A random number between 1 and 1000 is generated. If the number is larger than the crossover rate then reading continues with the second bit of A being entered into newA, and the second bit of B into newB and again checking for a random number less than the crossover rate. If a random number less than the selected crossover rate is encountered (after bit 5 for example), then the contents of newA and newB are switched at this point, and filling newA from A and newB from B continues as before. If this is the only crossover detected then the newA and newB will now be:

newA 0,1,1,0,0,1,0,0,1,0,1,0,1,1,1,..........1
newB 1,0,0,0,1,0,1,0,1,1,1,0,0,0,0,..........0

Exchanging string segments in this manner breeds new designs using the best partial solutions from the previous generation.

When a bit mutates, it changes (1 to 0) or (0 to 1). To accelerate the GA process the mutation operator is combined with the crossover operator into a single breed function. As each bit of strings A and B are examined, a mutation is implemented if a second random number (between 1 and 1000) is less than the mutation rate. Mutations provide a means for exploring local regions of the gene hyperspace in the vicinity of the fittest faces. Following selection, crossover and mutation, the new generation of faces is developed from the new genotypes and rated by the witness as before. This procedure continues until a satisfactory composite of the culprit has been evolved.

The genetic algorithm provides a unique search strategy that quickly finds the most "fit" outcome from a choice of evolutionary paths through the "face space". The strength of the algorithm lies in (a) the implicit parallelism of the search along all relevant dimensions of the problem (b) the exponential increase in any partial solution which is above average fitness, over generations, and (c) the exploration of small variations around partial solutions (Goldberg, 1989). Beginning with a set of points that are randomly distributed throughout the hyperspace, the selection procedure causes the points to migrate through the space, over generations. The efficiency of the migration is greatly enhanced by the sharing of partial solutions (i.e., best points along a particular dimensions) through crossover, and the continual exploration of small variations of these partial solutions, using mutations. The result is not a "random walk" but rather, it is a highly directed, efficient, and effective search algorithm whose success is attested to by its unrivaled role in biological adaptation (Dawkins, 1988).

(E) RESEARCH PLAN:

The proposed research program can be divided into (1) a design phase and (2) a testing phase. The goals of each phase are described below.

(1) Design Phase:

During the design phase, the current FacePrints program will be improved and modified in order to meet the requirements necessary for the testing phase. There are several proposed modifications expected to enhance the performance of the current algorithm.

(i) First, the spinning wheel selection operator, described above, will be replaced by a Stochastic Universal Sampling (SUS) procedure (Baker, 1987). The latter process involves a single spin using a number of equally spaced pointers corresponding to the generation size. SUS has been shown to reduce bias in the sample, thus selecting breeders more exactly in proportion to relative fitness.

(ii) A second modification to the GA involves the use of Gray code rather than binary.

Integer	Binary	Gray
0	000	000
1	001	001
2	010	011
3	011	010
4	100	110
5	101	111
6	110	101
7	111	100

The purpose of this is to allow single mutations to explore regions of the sample space adjacent to a fit phenotype. In biological systems, where genes are codes for proteins (e.g. enzymes), most mutations have minimal effects on the efficiency of the protein. Only the active sites, a small fraction of an enzyme's structure, is essential for activity. As a consequence, mutations which affect non-critical portions of the molecule often have little or no effect on enzyme performance. This mechanism allows mutations to "fine-tune" their phenotypic effects, rather than always producing radical changes. In Gray code, unlike binary, a single mutation is all that is required to move from any integer value to the next higher or lower decimal equivalent. This ability to "fine-tune" the phenotype by single mutations is particularly important when an "almost correct phenotype" has been evolved. For this reason, is expected that Gray code will be superior to binary.

(iii) A third improvement to the current GA, involves determining the optimal mutation and cross-over rates in order to speed up the search process. A meta-level GA will be used to evolve optimal values for these parameters.

(iv) The final goal of the test phase will involve an evaluation of the user interface during a pilot study. This pilot study will enable the experimenters to examine a number of design options which may contribute to the speed or ease of use of the FacePrints process.

(2) Testing Phase:

The testing phase is designed to evaluate the FacePrints process enhanced by all of the improvements introduced as a result of the design phase. "Witnesses" will be exposed to a simulated crime (on videotape) and then be required to make a composite of the culprit using the FacePrints program. The quality of the final composite will be determined by measuring the ability of judges to select the culprit from an array of faces, using the composite generated by the witnesses. Two variables will be examined for their effect on the performance of the program: the distinctiveness of the culprit and the delay between exposure to the videotape and generating the composite using FacePrints.

(F) RESULTS:

(1) Design Phase:

A simulated witness program (SAM) was developed to facilitate the development and testing of the proposed modifications. SAM was designed to simulate a "perfect" witness who could accurately rate each generated facial composite according to its phenotypic distance (resemblance) from the culprit. SAM made it possible to evaluate each design modification of Faceprints over hundreds of experimental runs.

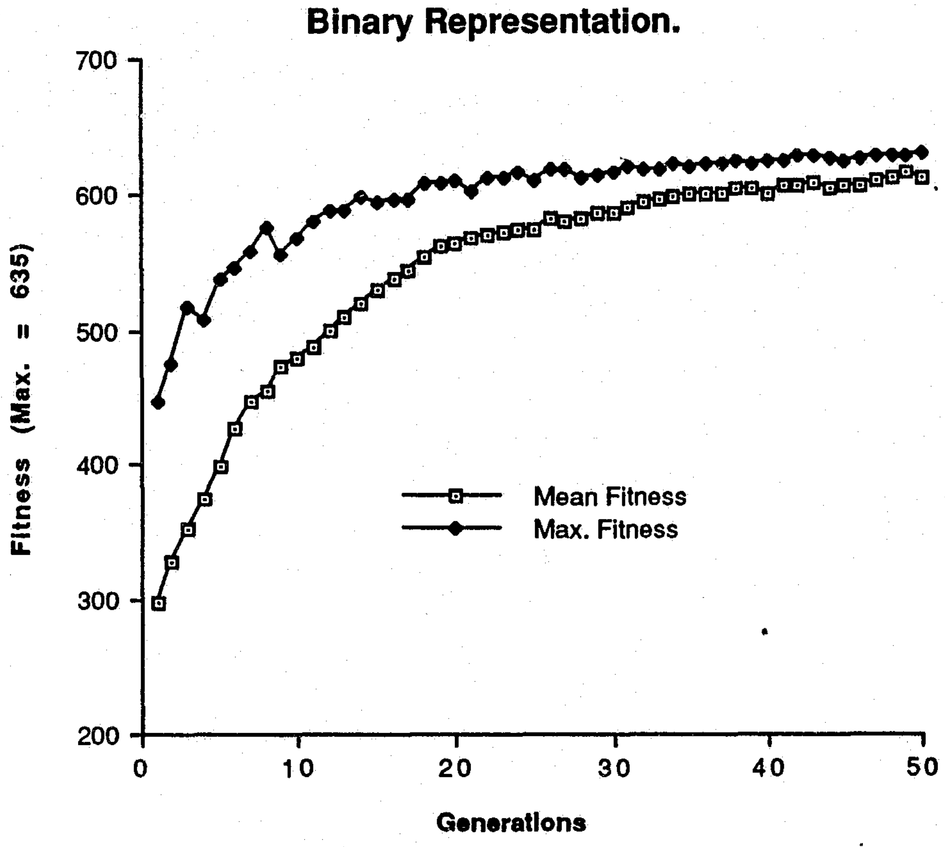


Figure 2: Increase in the fitness of a composite over generations, for a "perfect" witness (SAM).

(a) Stochastic Universal Sampling. The SUS procedure was incorporated into the GA. Figure 2 shows the improvement of a facial composite (fitness) over 50 generations, using SUS and the simulated witness. Both the average fitness of the population (mean fitness) and the fitness of the best composite (max. fitness) are shown. There were 20 composites per generation. SAM had evaluated only 400 composites out of 34,359,738,368 possible composites (.0001%), by generation 20. A real witness would require about 1 hour to make these evaluations. Since one hour and 400 evaluations are about the maximum values we could reasonably expect from a real witness, the performance after 20 generations (G20 performance) has been used as a benchmark. In Figure 2, the maximum possible fitness (perfect composite) is 635. The mean G20 performance is therefore 560/635 ; 88% of the maximum possible fitness: the best G20 performance is 610/635; 96% of maximum

	BINARY	GRAY	"BIN/GRAY"
0	000 000	000 000	000 000
1	000 001	000 001	000 001
2	000 010	000 011	000 011
3	000 011	000 010	000 010
4	000 100	000 110	000 110
5	000 101	000 111	000 111
6	000 110	000 101	000 101
7	000 111	000 100	000 100
8	001 000	001 100	001 000
9	001 001	001 101	001 001
10	001 010	001 111	001 011
11	001 011	001 110	001 010
12	001 100	001 010	001 110
13	001 101	001 011	001 111
14	001 110	001 001	001 101
15	001 111	001 000	001 100
16	010 000	011 000	010 000
17	010 001	011 001	010 001
18	010 010	011 011	010 011
19	010 011	011 010	010 010
20	010 100	011 110	010 110
21	010 101	011 111	010 111
22	010 110	011 101	010 101
23	010 111	011 100	010 100
24	011 000	010 100	011 000
25	011 001	010 101	011 001
26	011 010	010 111	011 010
27	011 011	010 110	011 110
28	011 100	010 010	011 111
29	011 101	010 011	011 101
30	011 110	010 001	011 100
31	011 111	010 000	011 000

Figure 3: Binary, Gray and "BinGray" codes.

(b) Gray Code and Binary Code Evaluation (Figure 3). A potential problem with binary code can be seen when moving from decimal 3 (binary 011) to decimal 4 (binary 100). If decimal 4 is a fit phenotype then decimal 3 also has high fitness. However, at the level of the genotype (binary) it requires three simultaneous bit mutations to move from 011 to 100. This "hanging cliff" can be avoided by using Gray code, where a single mutation is all that is required to move from any integer value to the next higher or lower decimal equivalent.

Figure 4, shows the effects of Binary and Gray code on GA performance using SAM. Over multiple simulations, the G20 performance of binary (88.6%) was always superior to Gray (81.1%). The problem with Gray appears to be the inconsistent interpretation of the most significant bits. A new code (BinGray in Fig 3) was tested, which uses binary for the most significant bits and Gray for the least significant bits. The average G20 performance of BinGray was 87.1%. We have therefore returned to binary coded genotypes. (We believe that further studies of other possible codes, and how they interact with the mutation and crossover operators, would be valuable for improving the performance of the GA).

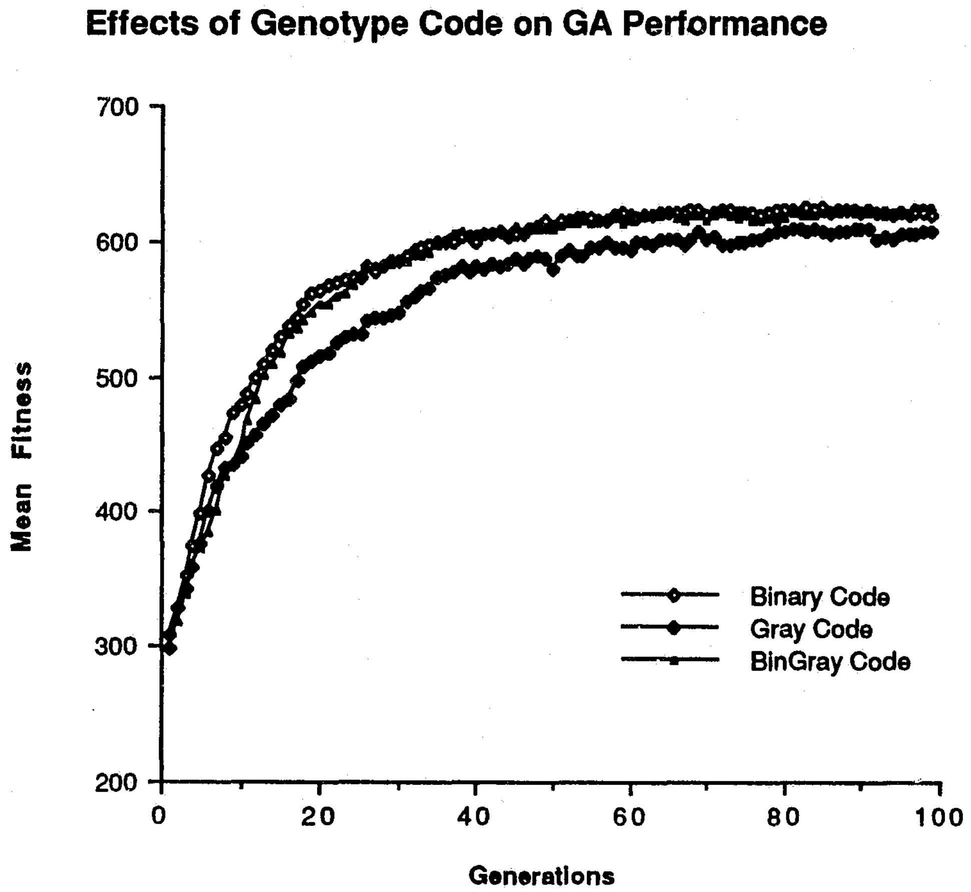


Figure 4: The effects of Binary, Gray and BinGray on the performance of the GA.

(c) Optimizing the GA parameters. The most significant advance in this phase of the experiment was the development of a method for determining the optimal crossover and mutation rates for use with Faceprints. This procedure involved the use of a second GA (meta-level GA), where binary, meta-level strings, coded for the crossover and mutation rates. Each string was then evaluated by determining how well the simulated operator (SAM) evolved a composite face using the crossover and mutation rates specified by this string; the G20 performance was used to measure the fitness of the meta level string. The meta-level population was then evolved, over a series of generations, to breed the optimal rates. This meta-level GA has been used in sequentially improved versions of Faceprints.

(d) Evaluating the user interface: The aims of this phase were to evaluate the user interface of the FacePrints program, determine if the current implementation of FacePrints was sufficiently fast for practical use, and examine the effects of presenting the best facial composite from the previous generation to a witness rating the current generation.

Subjects. The subjects were 40 undergraduate student volunteers (20 male and 20 female) who were randomly assigned to the two experimental groups. Subjects in Group N (10 males and 10 females) did not have the best composite from the previous generation available when rating the current generation. Group C (10 males and 10 females) subjects were provided with their best prior composite (displayed on the right side of the computer screen) while rating the current generation of faces.

Apparatus. FacePrints (version 1.0 - HyperTalk version) was used in this experiment. This version generated an initial set of twenty random bit strings (genotypes), each string being 35 bits long. The 20 strings were decoded into facial composites (phenotypes) using seven bits to specify each facial feature (hair, eyes, nose, mouth and chin) and its position relative to a fixed eye line. (The eye position bits specified the distance between the eyes). The 35 bit code had the potential to generate over 34 billion unique composites.

Each new generation of faces was produced by breeding together the highest rated faces from the previous generation, using Stochastic Universal Sampling and the optimal crossover and mutation parameters derived from the meta-level analysis.

Procedure. Each subject was exposed to a ten second display of a standard culprit's face. Immediately following this display they used the FacePrints program to evolve a facial composite of the culprit. The subjects were told to (a) rate each of the first generation of faces, on a 9 point scale (fitness), according to its perceived resemblance to the culprit, (b) wait until a new generation of faces was generated by the computer program, (c) rate each of the new generation of faces, and (d) repeat steps (b) and (c) until a one hour experimental session was completed. The subjects in Group C were informed that after the first generation the most highly rated composite from the previous generation would be displayed as a reference, while they viewed the new generation of faces. They were instructed to consider this composite as having a value of 5 on the 9 point rating scale, and to rate a current composite as higher than a 5 only if it had a closer resemblance to the culprit than this reference face.

Results. There was a wide variation in performance between subjects. The number of generations completed within the one hour session varied from 7 to 12. For the purpose of data analysis, the generation 7 composite (G7) was examined for all 40 subjects. Two measures of the quality of G7 were used; a "subjective" and an "objective" measure.

The "subjective" measure was obtained by having 12 naive raters (6 male and 6 female) examine the G7 composites of all 40 subjects and rank them for degree of resemblance to the culprit. An analysis of G7 composites revealed no significant difference in quality between the two treatment groups.

The "objective" measure of quality was computed as the phenotypic distance, in the data base, of the G7 composite from the culprit. That is, (hair distance + eye distance + nose distance + mouth distance + chin distance) divided by 5. If the G7 hair was correct, then the hair distance would be zero; if the G7 hair was one above or below the culprit's hair in the data base order, then the hair distance would be 1. (This phenotypic distance is the same measure used by the simulated witness as discussed in the previous report).

Figure 5a shows the change in phenotypic distance, over generations, for the two experimental groups. Although the rate of improvement over generations appears greater for the subjects in Group C, there was no significant difference between the treatment groups in terms of G7 performance. Figures 5b, 5c, 5d, 5e and 5f show the change in the phenotypic distance from the culprit for Hair, Eyes, Nose, Mouth and Chin respectively, over the 7 generations, for both groups of subjects.

Discussion. The purpose of the pilot study was (a) to evaluate the gains or losses associated with presenting the prior generation best composite during the rating of the current generation and (b) to test the user interface of the FacePrints program.

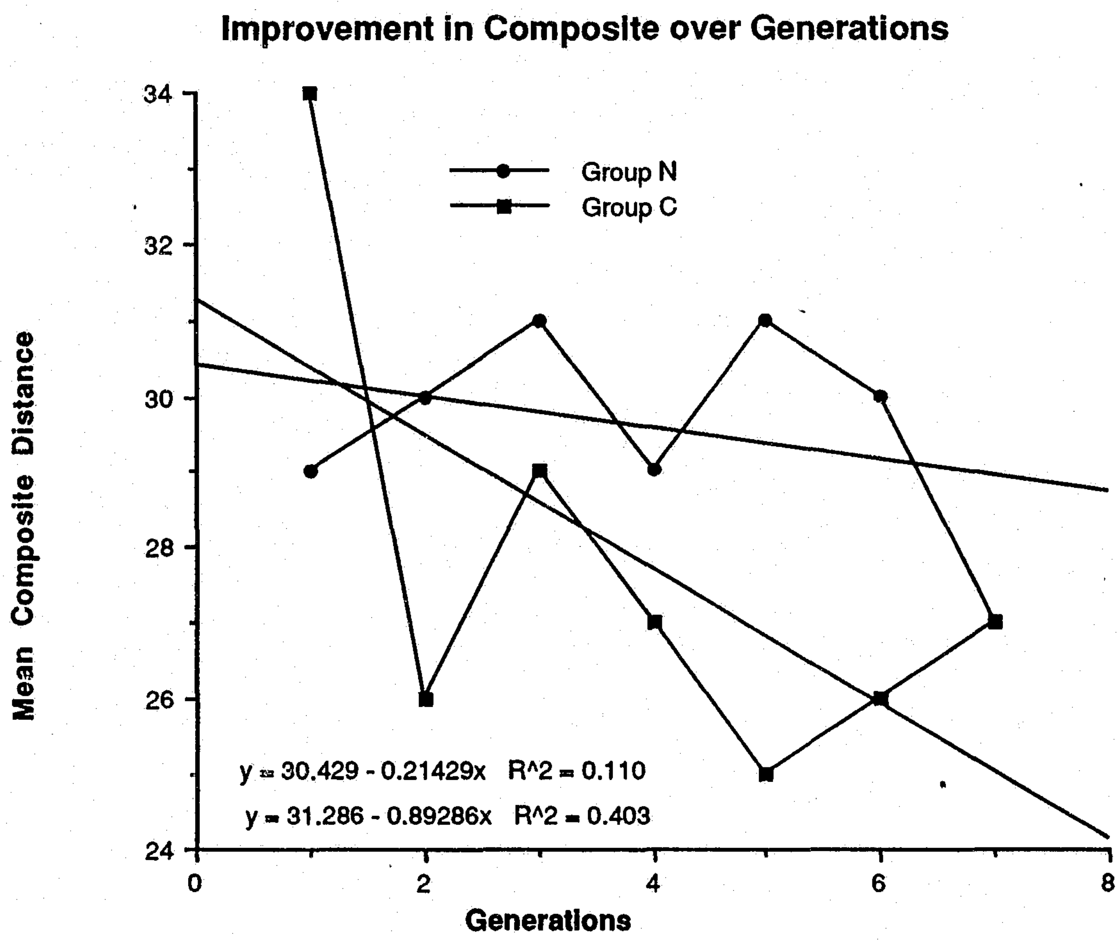


Figure 5A

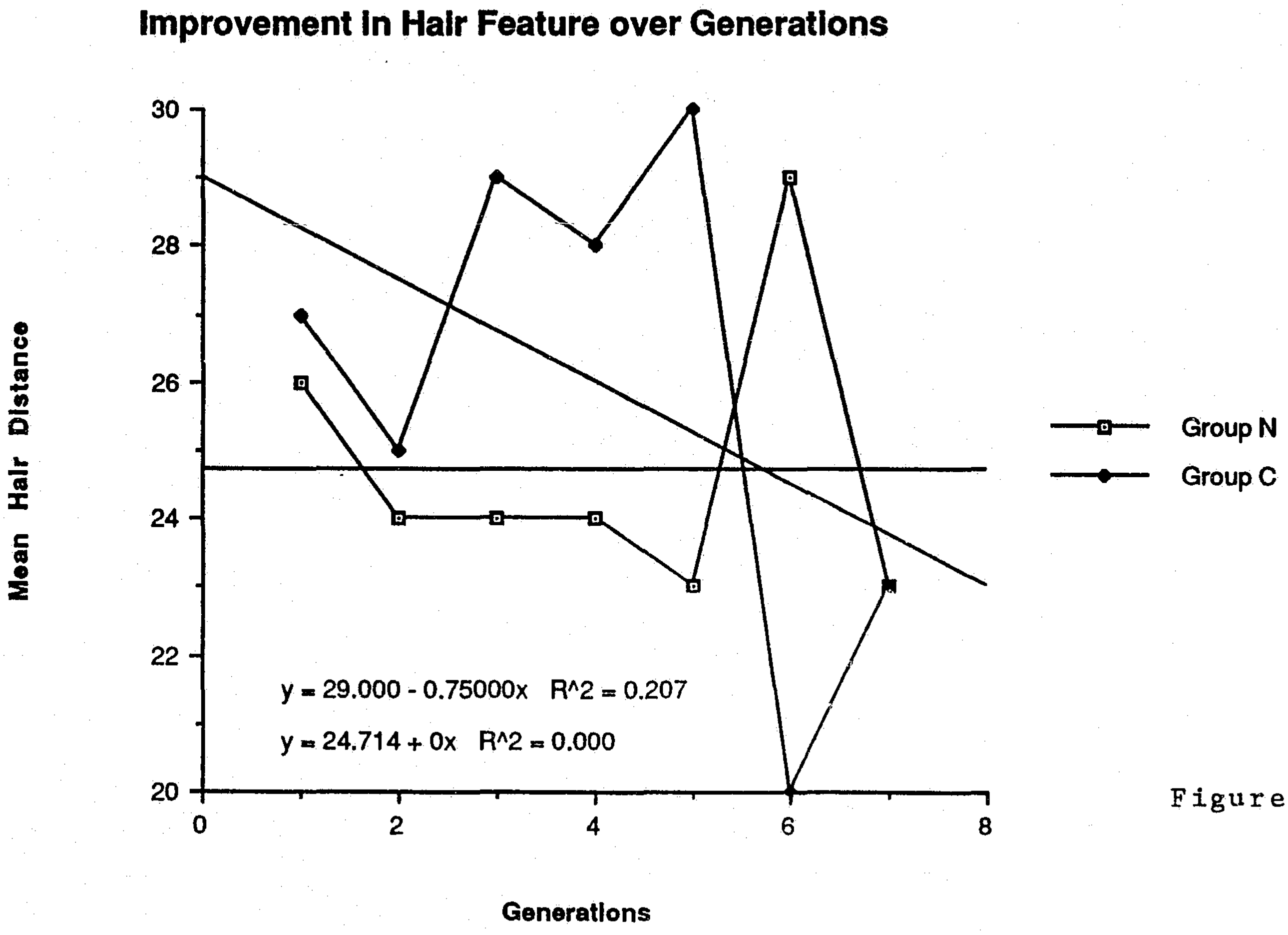


Figure 5B

Improvement in Eye Feature over Generations.

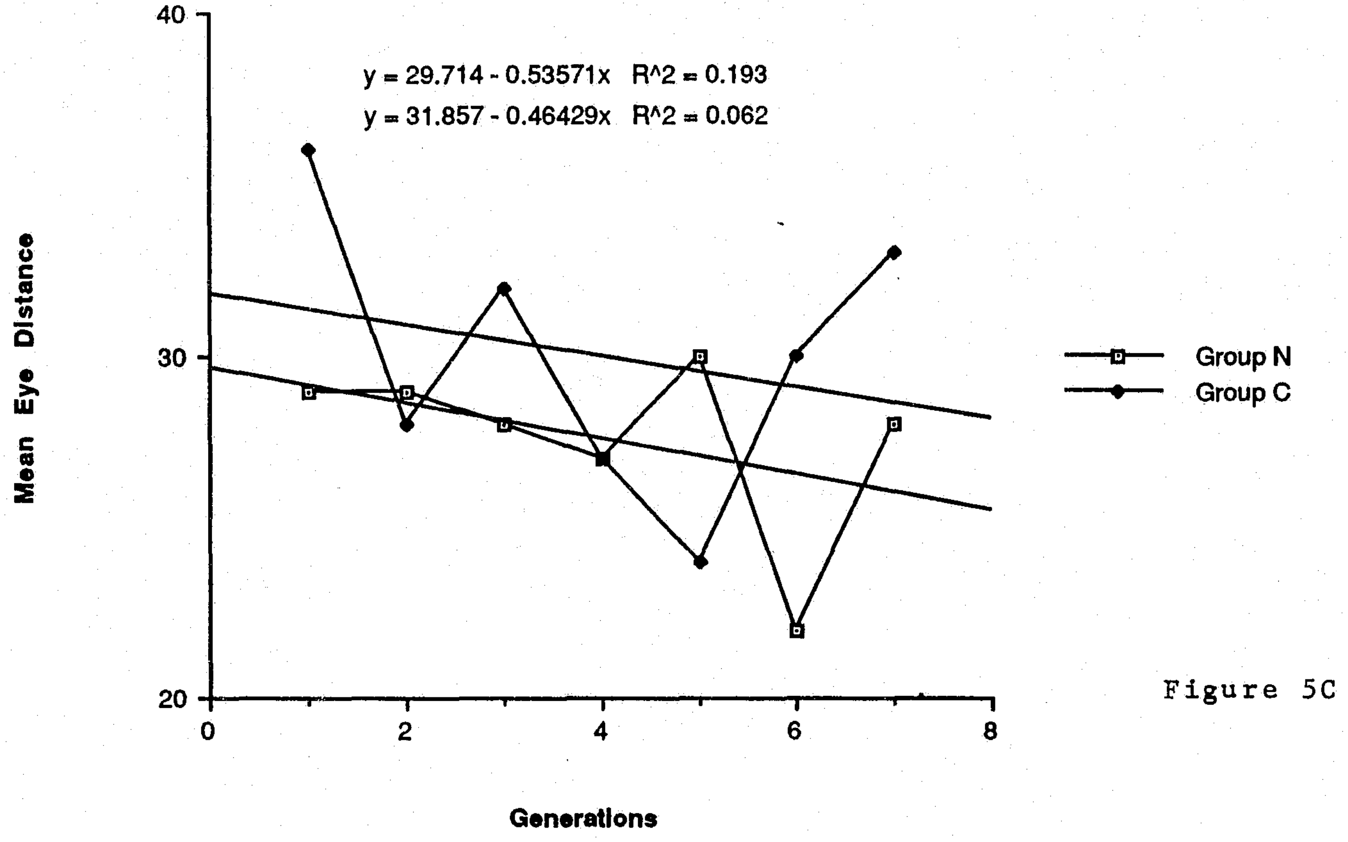


Figure 5C

Improvement in Nose Feature over Generations

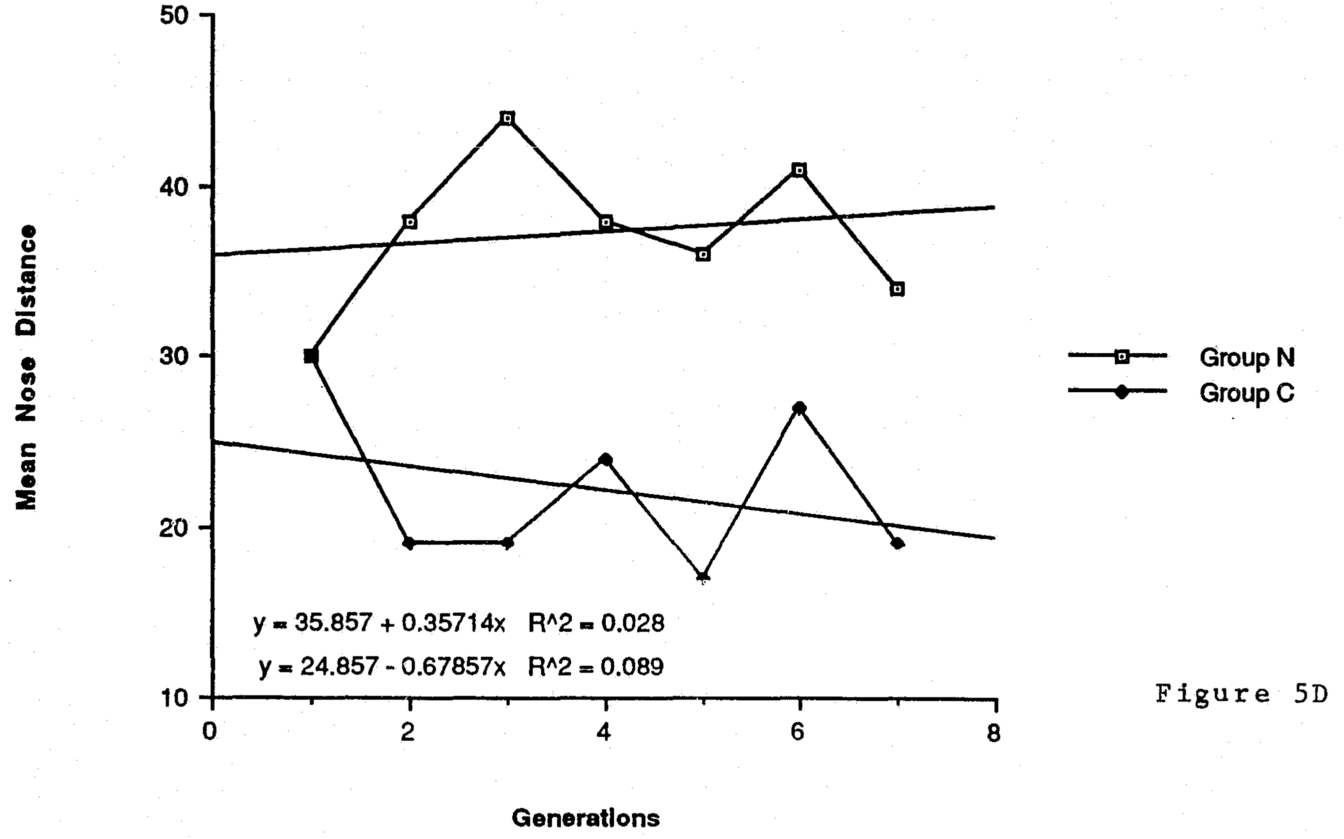


Figure 5D

Improvement in Mouth Feature over Generations

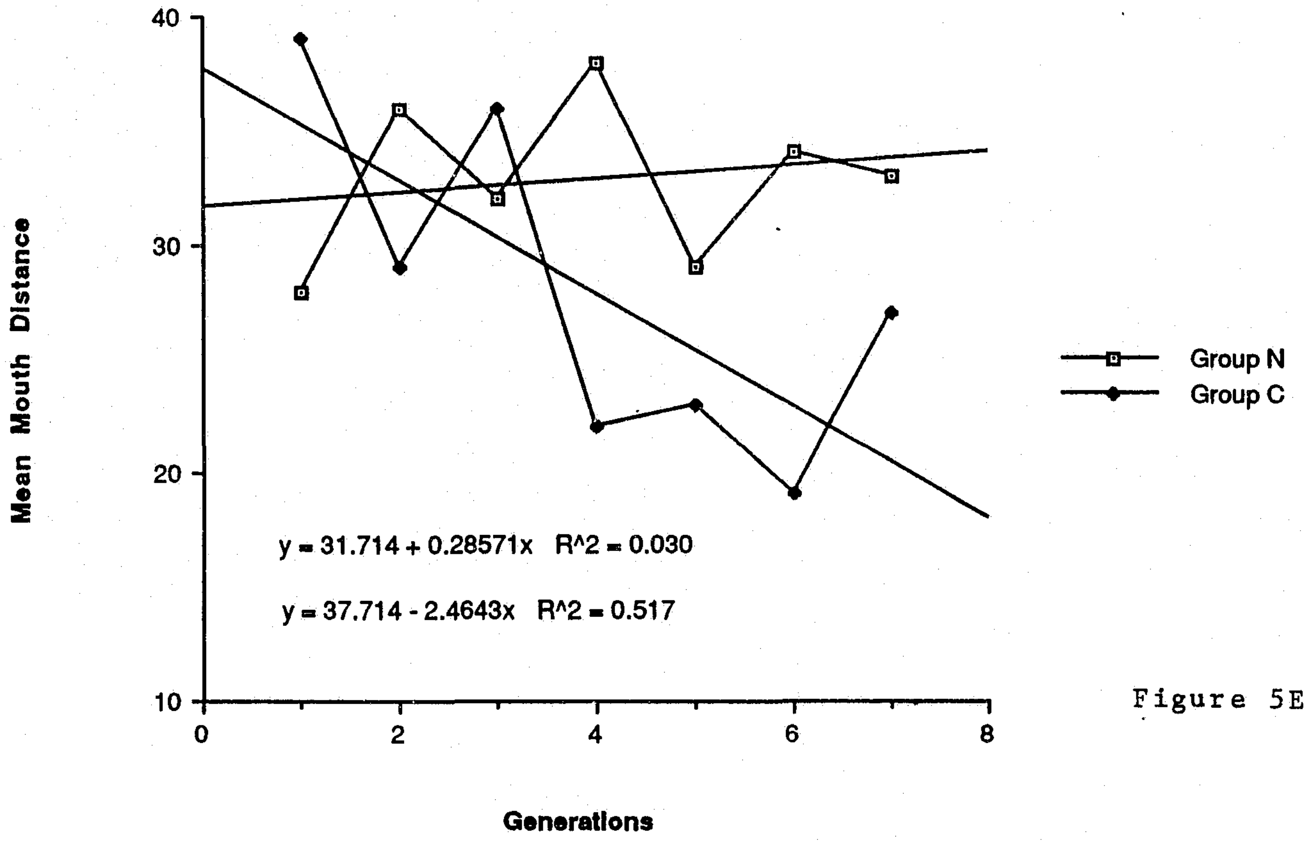


Figure 5E

Improvement in Chin Feature over Generations

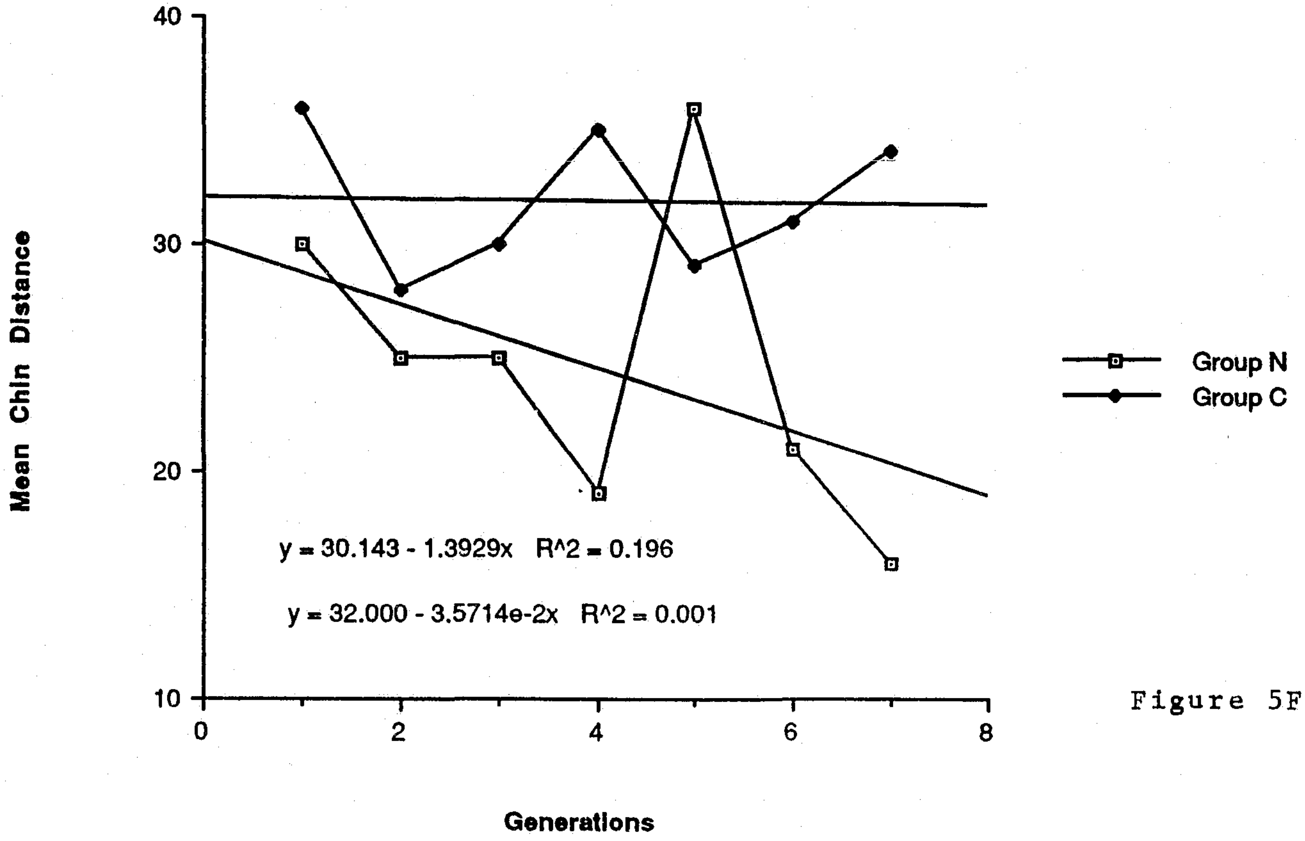


Figure 5F

No significant differences in the quality of the final G7 composite were obtained using the two experimental procedures. However, Figures 5b to 5f reveal that subjects using a reference composite did show a more systematic improvement in all features over generations; all regression line slopes are negative in value. This parallel improvement in all features is the major strength of the FacePrints procedure. It is also clear from the slope of the regression lines that some features (e.g. chin-slope = - 1.39) were being selected more than other features (e.g. nose-slope = - 0.67). This suggests that some facial features may be generally more important than others in determining the degree of similarity between two faces. Haig (1986a) has also noted that the head shape (hair + chin) is the dominant feature used for recognition, and that the nose is the least significant feature. It is clear that an expanded study using the FacePrints program could provide quantitative data on the relative importance of facial features and cephalometric measurements for recognition. This data would be of great value in aiding the design of any facial composite system since it provides insights into the size of the data base needed for each feature.

The user interface was satisfactory, with the following exceptions. Some subjects found it difficulty to use the computer mouse to click on the rating scale Consequently, keyboard function keys (F1 to F9) were implemented as an alternative way to input ratings. In addition, subjects were frustrated by the delay between generations (almost 3 minutes) and the inability to "keep" what they perceived to be a good composite. They often complained that good features were lost between generations! The next section outlines the modifications to the FacePrints program in order to overcome these difficulties.

(e) Program Modifications: FacePrints (version 2.0) was subsequently rewritten in SuperTalk, a commercial application program designed as an improvement to HyperTalk. Implementation in SuperTalk reduced the inter-generation time from 3 minutes to 18 seconds. At the same time, the computer interface was redesigned to permit keyboard inputs for all operator controls. Audio signals and flashing buttons were added to prompt the user in the event of a long delay in any input response.

Based on the pilot study findings, the best composite from the prior generation was concurrently displayed while subjects rated the composites of each successive generation. Comments from the subjects on the use of the prior composite suggested additional options which could enhance the effectiveness of the FacePrints process and, at the same time, overcome the subjects' "frustration" in the loss of good features between generations.

Flood Option: When subjects rated any generation of (20) composites, the highest rated composite from that generation was displayed in a window of the computer screen. Before breeding the next generation, subjects were now permitted to lock one or more features of this composite (hair, eyes, nose, mouth or chin). That section of the 35 bit string corresponding to the locked feature was then inserted into all the genotypes of that generation, before breeding. Since all genotypes were then identical at the location of the locked feature, the cross-over operator could not modify that feature in the next generation of faces. (There is still a small probability of modification by mutation.)

Freeze Option: A variation of the above procedure, the Freeze option, was implemented in a similar manner, but now the locked feature was also protected from mutations.

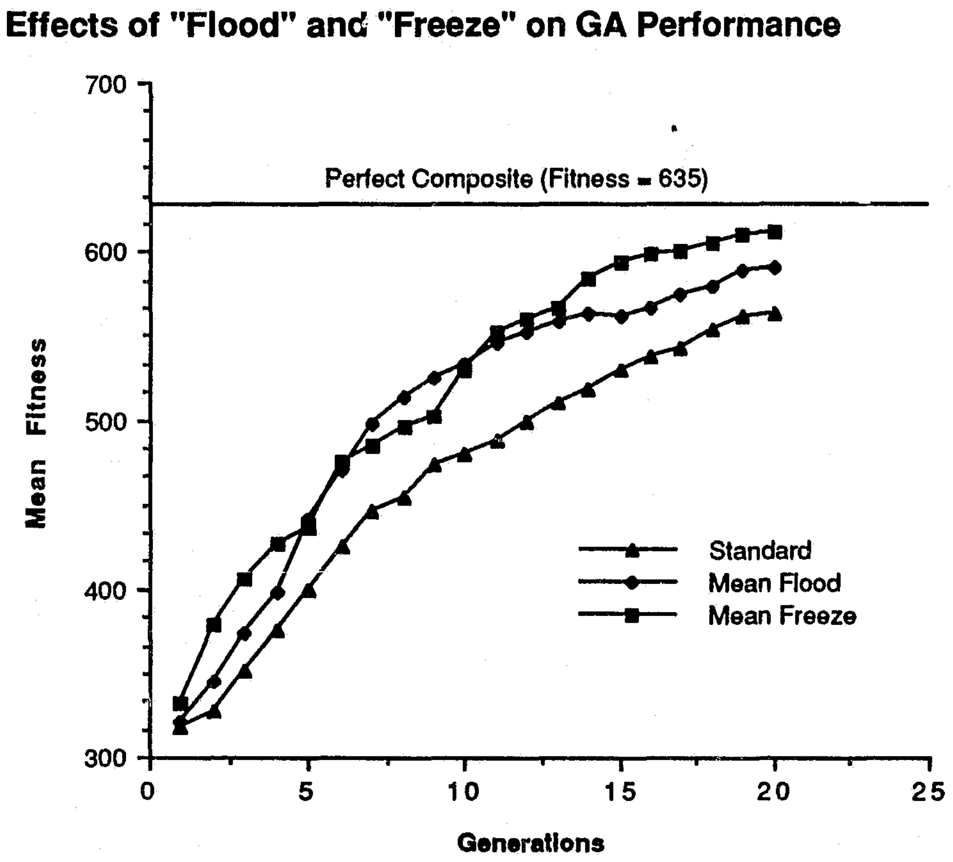


Figure 6: Improvement in fitness over generations when "perfect witness" (SAM) uses Freeze or Flood option.

(f) Evaluating Flood and Freeze Options: In order to evaluate both of these locking procedures, it was first necessary to evolve the optimal cross-over and mutation parameters for each technique. Results obtained from running the meta-level program (see above) revealed that the optimal probability of a crossover was 0.24 for both options, but the optimal mutation probability was 0.03 for the Flood option and 0.05 for the Freeze option. The simulated operator program (SAM) was used to compare the expected performance of FacePrints with and without these two options. Figure 6 shows the results of this analysis. The G20 performance (refer to prior report) revealed that both the Freeze and Flood options produced a marked improvement in the performance of the algorithm (Standard G20 = 88.6%, Flood G20 = 93,1%, Freeze G20 = 96.4%). The superior performance of Freezing over Flooding probably resulted from the harmful effects of mutations as composites began to converge to the likeness of the culprit. Mutations in early generations may have enhanced performance, (by exploring more areas of the data base) but in later generations these mutations have a higher probability of being destructive.

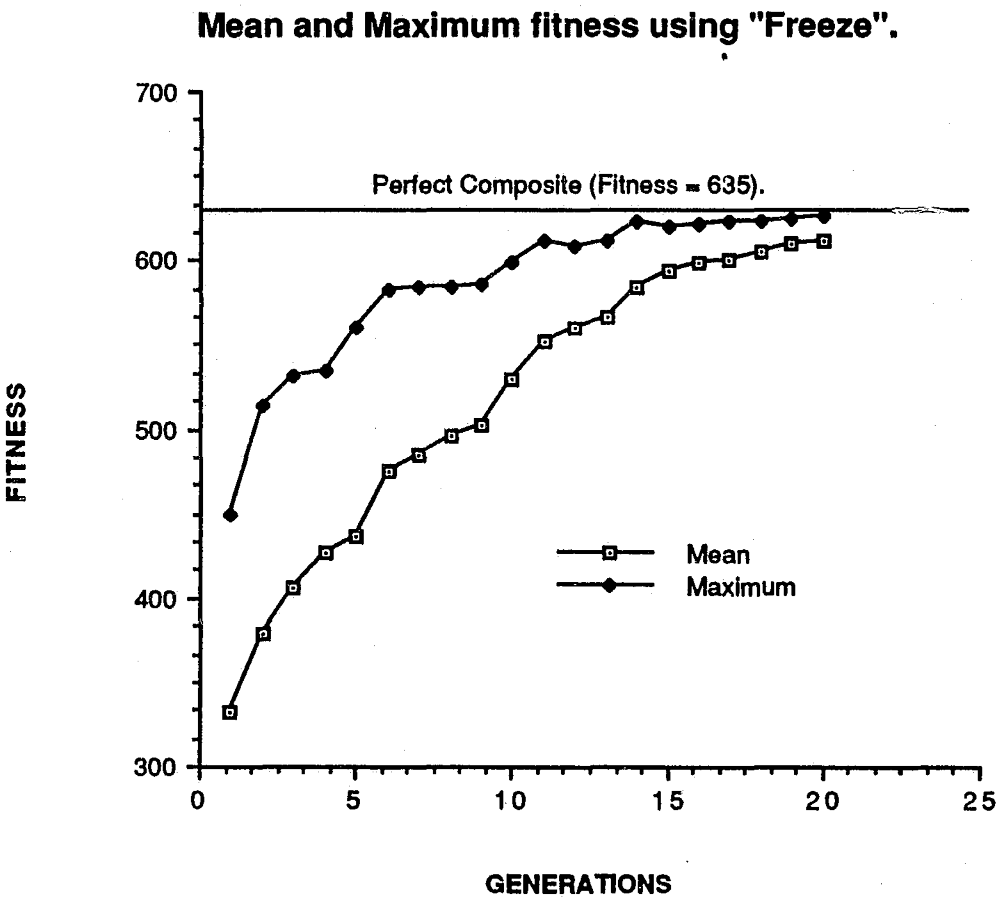


Figure 7: Improvement in the fitness, over generations, for a "perfect" witness (SAM) who can freeze features.

Figure 7 shows the mean fitness of the population and the fittest composite within each generation (maximum), using the Freeze option. The mean and maximum performance at generation 10 and 20 were: G10 = 83.3, 94.3; G20 = 96.4, 98.6, respectively. These results suggest that a substantial likeness to the culprit can be achieved after only 10 generations, if the behavior of a real witness approximates the behavior of the simulated witness. This is a very encouraging result since it establishes that in theory it is possible to find a single almost perfect composite (out of 34 billion) by rating only 200 composites (less than 0.000,000,6%). For this reason, the Freeze option has been included in FacePrints (v 3.0) for use in the major experiment.

(2) Testing Phase:

Three male volunteers were selected to act as culprits in this experiment. Each culprit performed a simulated armed robbery which was recorded by a "surveillance" camera. In the final act of the robbery the culprit's face could be seen as he turned to shoot at the camera. This segment of the video was frozen, so that a witness viewing the tape could see a 10 second still-frame of the culprit's face. This method of presentation ensured that all witnesses receive an equal exposure to the same view of the culprit. (See enclosed videotapes).

(a) Culprit Selection: The three culprits were selected to represent three levels of facial distinctiveness within the data base. Both MUG shots and "natural-with-expression" photographs were taken of each culprit, after the crime was staged. Culprit's features were scanned from the front face MUG shots and introduced into the data base. Figure 1 shows the MUG shots of each culprit.

The first culprit (Figure 8, top) was selected because his features and cephalometric measurements were similar to the average face. It is difficult to provide a precise quantification of the degree of similarity, but three different measures support this claim. First, for each culprit, the distance of each feature from the center of the data base was determined. The sum of these distances can be viewed as an approximate measure of feature distinctiveness. Using this scale, the first culprit scored 28, with culprits 2 and 3 scoring 35 and 31 respectively. The problems with such measurements are:(1) they depend on the organization of the data base, which is only approximate since each feature is multidimensional (e.g mouth width, thickness of top lip, thickness of bottom lip, etc.) (2) all features receive equal weight - an unjustified assumption, and (3) the size of the jumps between the elements in the data base is an arbitrary distance.

A second approach was to compare each culprit's face to the average face in the population. For this measurement, the average was considered

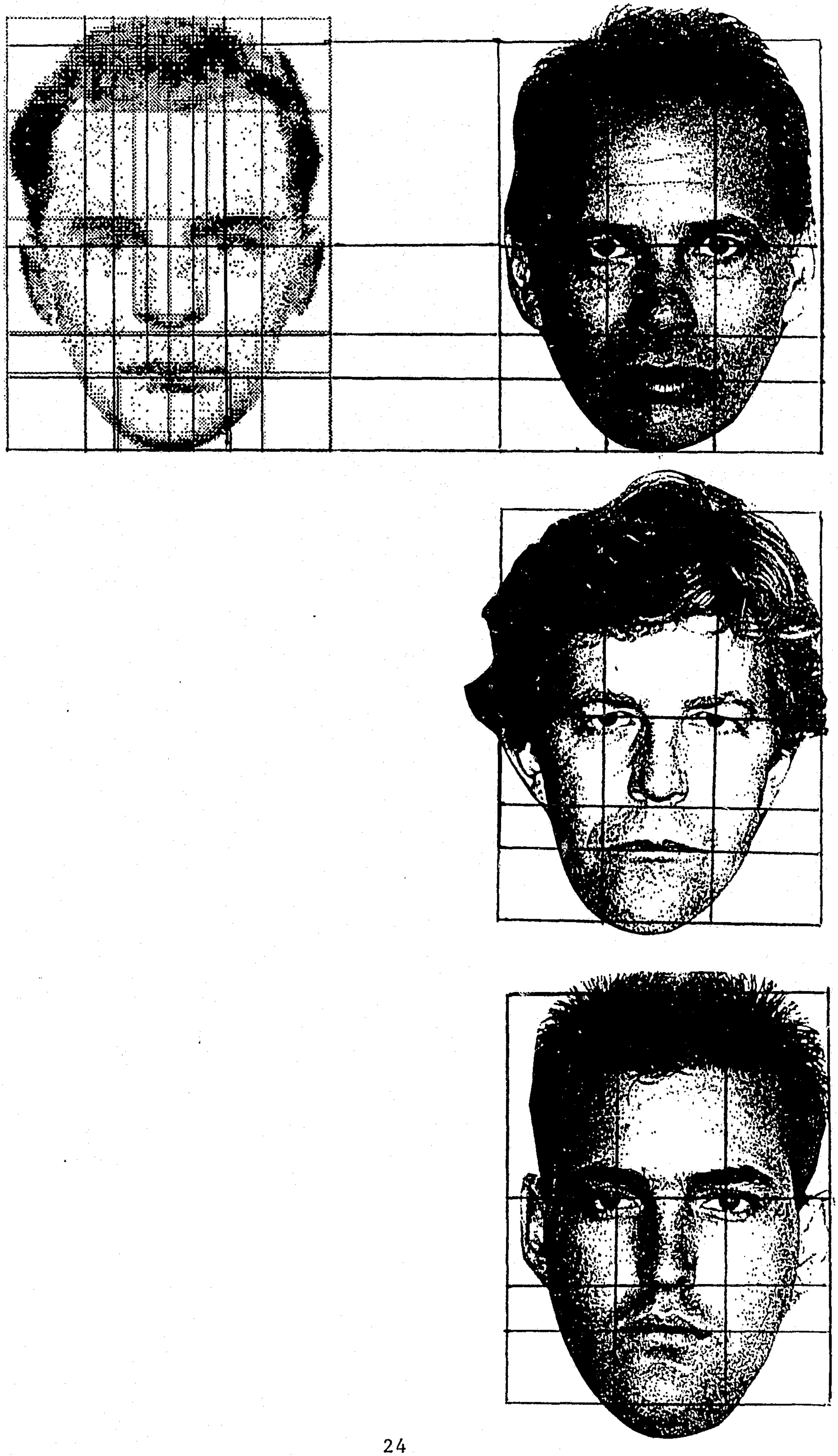


Figure 8

to be the average face shown in the PhotoFit Manual, and reproduced in Figure 8. The distance of each culprit's features (hair, nose, mouth, chin) from the average eye line, and the distance between the eyes compared with the average eye distance, could then be measured to provide a composite score which approximates each culprit's distinctiveness with respect to average cephalometric measures. On this scale, the three culprit's scored 3, 6, and 5 respectively, with culprit 1 measuring closest to the average face. Again, the feature distances are not weighted for importance, and it is impossible to justify equating a one millimeter change in eye distance with a one millimeter displacement of the chin.

The third and perhaps most valid determination of distinctiveness was obtained by requiring a group of twenty judges to rank the culprits' faces with respect to their similarity to the average face. Twelve subjects (60%) rated culprit 1 as closest to the prototype, with culprits 2 and 3 receiving 15% and 25% of the vote, respectively.

All three measures suggest that culprit 1 has features which are closest to a prototypical face, with culprit 2 possessing the most distinctive features and cephalometric measurements.

Prior to the main experiment, two pilot studies was conducted in order to determine (1) the degree of difficulty in recognizing a culprit from a videotape exposure, and (2) the degree of difficulty in recognizing a culprit from a bit-mapped composite face of that culprit.

(b) Pilot Study 2: In this study 45 student volunteers (15 males, 30 females) were randomly assigned to three experimental groups, with 15 subjects in each group. Each subject viewed one of the videotapes of the simulated robbery; a different culprit was used for subjects belonging to each of the experimental groups. Subjects were immediately required to select the culprit from a display of 36 faces. These faces included the "natural-with-expression" photographs of the target subjects and a selection of "natural-with-expression" photographs taken from the general student population (Figs 10a to 10d). The degree of recognition was determined by requiring subjects to make 5 selections from the 36 faces. Subjects were considered to have good recognition ability if they selected the target face on their first choice, and some recognition if the target was within their first five selections.

Results: Figure 9 shows the recognition performance of subjects as a function of the target face. Recognition performance was influenced by the nature of the target ($X^2{}_2 = 9.26$; $p < .01$), with recognition being worst for culprit 1, and best for culprit 3. The recognition of culprit 1 was significantly poorer than recognition of the other culprits ($X^2{}_1 = 13.26$; p <

.001). For the three culprits, the probability of recognition following videotape exposure was 0.33, 0.66 and 0.87, respectively.

An analysis of good recognition (selection of the target on first choice) also revealed a significant effect of the target ($X^2_2 = 6.34$; $p < .05$). Again, the recognition of culprit 1 was poor compared with the other two culprits ($X^2_1 = 5.46$; $p < .05$). The sex of the subject had no effect on any measure of recognition ability.

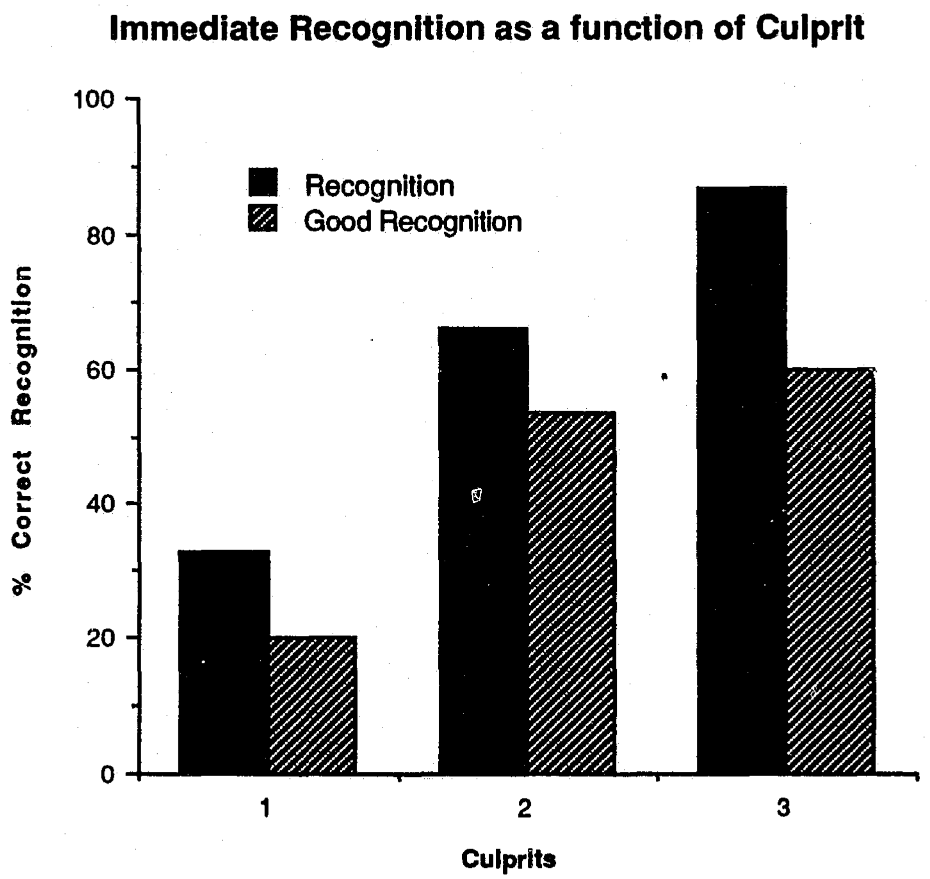


Figure 9: Recognition ability as a function of Culprit.

Discussion: It is apparent from this pilot study that the recognition of the culprit from the videotape is a difficult task. Subjects received a single brief exposure to the culprit's face. During this exposure the culprit's facial expression was quite different from that portrayed in the "natural-with-expression" photograph, shown during the recognition task. Furthermore, the alternative faces presented in the recognition set were similar to the culprit in age, racial origin, etc., and lacked any obviously distinctive features (e.g beards, glasses, etc.) which might have been an aid to recognition.

Recognition ability varied with the culprit. When the most prototypical culprit (Culprit 1) was the target, recognition was poor. This result supports the general finding that the more similar a target face is to the prototype, the less likely it is that the target face will be accurately recognized (Posner, 1973; Neumann, 1974; Solso and McCarthy, 1981; Light, Kayra-Stewart and Hollander, 1979; Valentine and Bruce, 1986; Haig, 1986). It is possible that some qualitative attribute of the videotape or natural photograph of Culprit 1 could account for the poor recognition, but no such quality differences are immediately apparent.

In view of these findings, we can conclude that the videotaped crimes present subjects with a difficult recognition task and that the selected culprits provide facial characteristics which vary in distance from a prototypical face. These videotapes and culprits are therefore suitable for evaluating the the FacePrints process for generating facial composites.

(c) Pilot Study 3: A second study was conducted in order to determine the probability of identifying the correct culprit in the recognition set when a subject was presented with a "perfect" bit-mapped composite of the culprit. In this study twenty subjects were shown bit-mapped composites (MUG shots) of all three culprits. Their task was to examine each composite and then identify that culprit from the 36 "natural-with-expression" photographs which made up the recognition set (Figures 10a, 10b, 10c and 10d). Identification was assessed by requiring the judges to make 5 selections from the 36 faces, in their order of preference. Judges were considered to have made an identification if the target face was within their five selections.

The probability of identifying the correct culprit from a bit-mapped composite varied with culprits. For culprit 1, the probability of correct identification was 0.45, whereas the other culprits were both identified with a probability of 0.7. These results indicate that the identification of a "natural-with-expression" photograph from a bit-mapped image is not a trivial task. As noted above, for recognition from videotape exposure, the alternative faces were similar to the culprits in age, racial origin etc., and they lacked any obviously distinctive features (e.g beards, glasses etc.) which might have been an aid to recognition. The observed failure to correctly identify the culprit from a "perfect" composite has importance for evaluating the FacePrints process (see later).

(d) Evaluation of the FacePrints process:

Subjects: One hundred and twenty one student volunteers (52 males and 69 females) served as subjects.

Apparatus: Three videotaped recordings of simulated crimes (described above) were used to expose subjects to a culprit's face.

Subjects were required to generate a composite of the culprit's face using the FacePrints process together with the freeze feature option (see previous reports) and the optimal cross-over and mutation rates as determined by the simulation program (see previous reports). The cross-over and mutation probabilities were 0.24 and 0.05 respectively. Each subject used the FacePrints process for ten generations, which required less than one hour of experimental time.

Recognition performance was assessed by requiring the subject to select the culprit's face from a selection of 36 faces This recognition set included the three "natural-with-expression" photographs of the target subjects together with 33 other "natural-with-expression" photographs from the general student population (Figs 10a, 10b, 10c and 10d).

Procedure: The 121 experimental subjects were randomly assigned to three delay groups. Subjects in group Delay 0 (19 males, 20 females) were individually exposed to one of the simulated crime videotapes and then required to immediately compose a facial composite of the culprit who appeared in that videotape. Each of the three culprits served as the target face for a randomly selected 13 subjects within·this delay group. Subjects belonging to group Delay 3 (17 males, 21 females) were treated in a similar manner, but a three day period was allowed to elapse between their exposure to the videotape and their production of the composite face. Within this group, the number of subjects using Culprit 1 as the target was 14, with Culprits 2 and 3 being the target for 14 and 10 subjects, respectively. Subjects in group Delay 7 (15 males and 29 females) were required to wait 7 days between exposure and production of the composite face. Culprits 1, 2, and 3 served as the target for 15, 14, and 15 subjects, respectively.

Immediately after generating a composite face, subjects in the Delay 3 and Delay 7 conditions were tested for target recognition. In the recognition task the subjects were required to select their target from the array of 36 "natural-with-expression" faces. Each subject was required to make five selections in order of preference. (The importance of determining recognition ability was not apparent until all Delay 0 subjects had completed their experimental task. This data is therefore not available for the Delay 0 subjects, but the performance of the pilot subjects provides a good estimate of recognition ability in the absence of a delay). Subjects were considered to have some recognition ability if the target face was within their five selections, and good recognition if the target was their first choice.

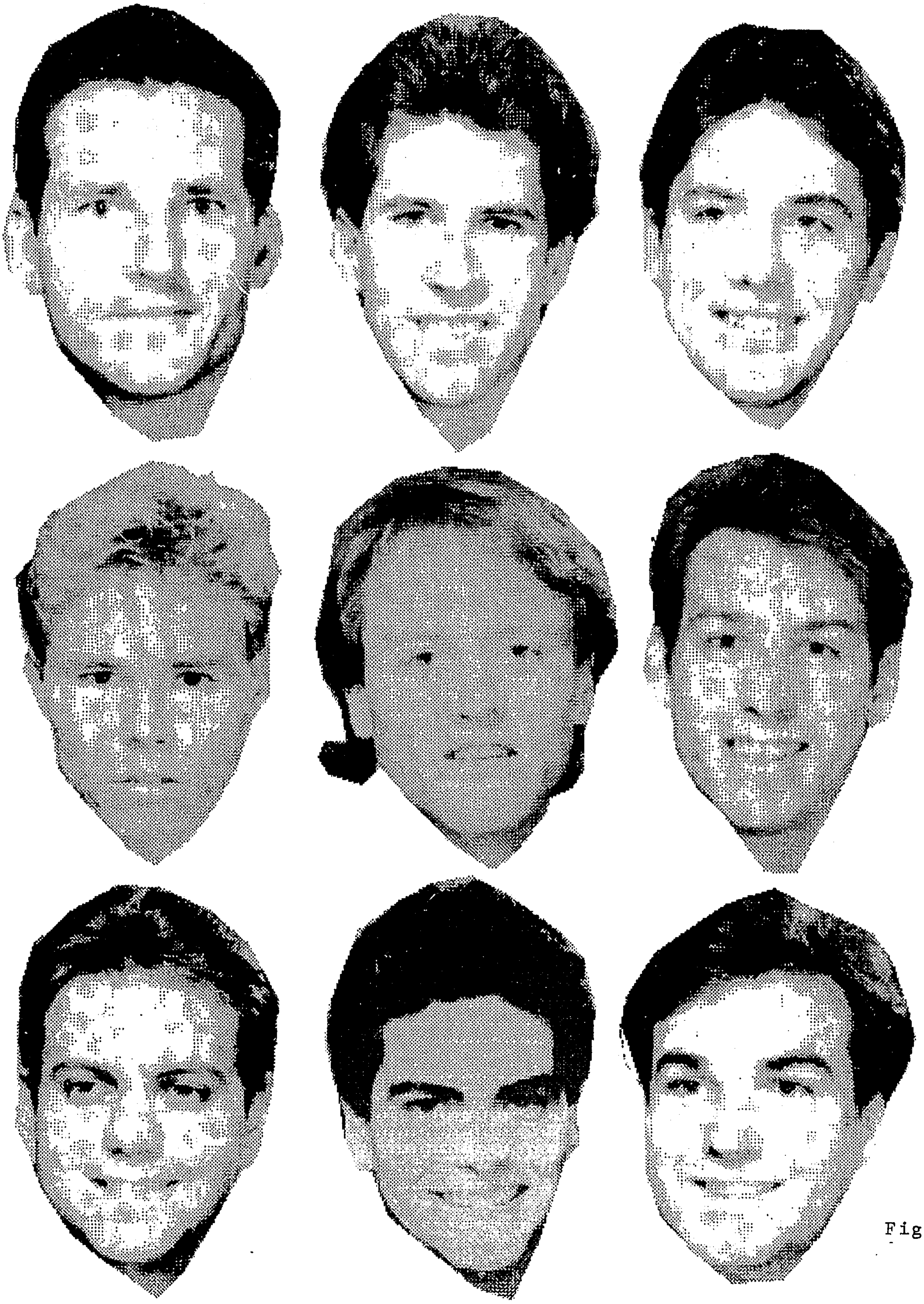


Figure 10A

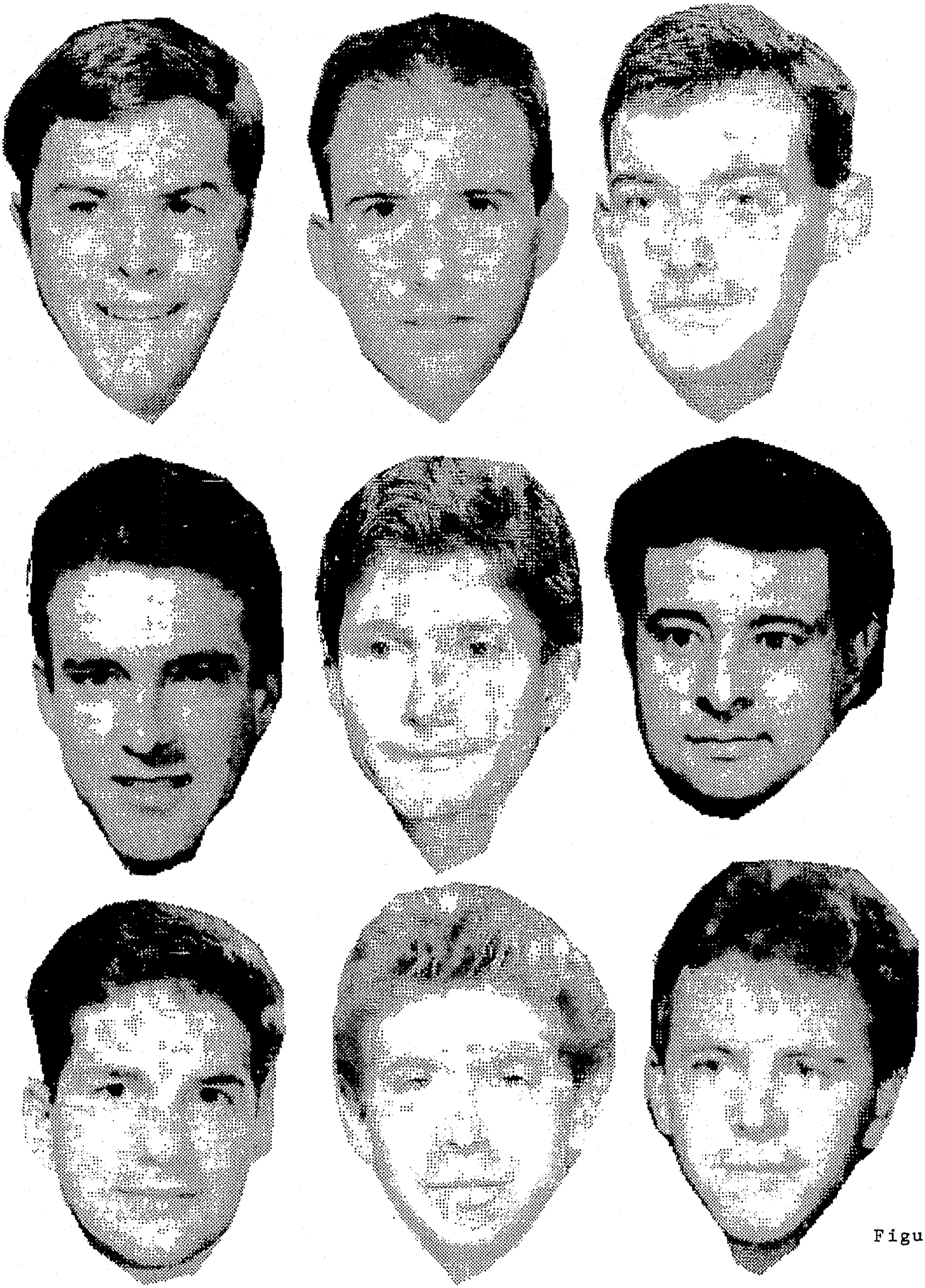


Figure 10B

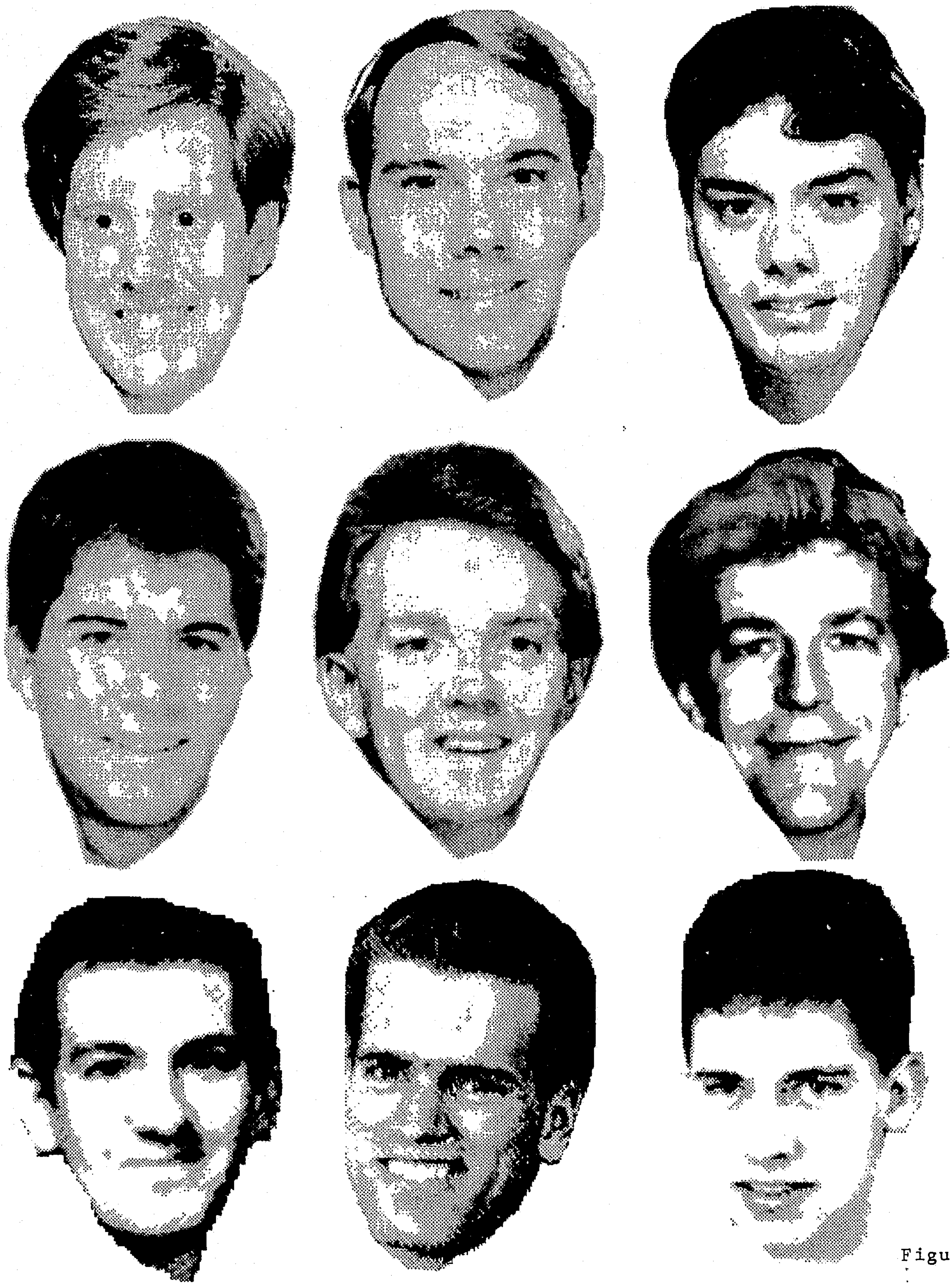


Figure 10C

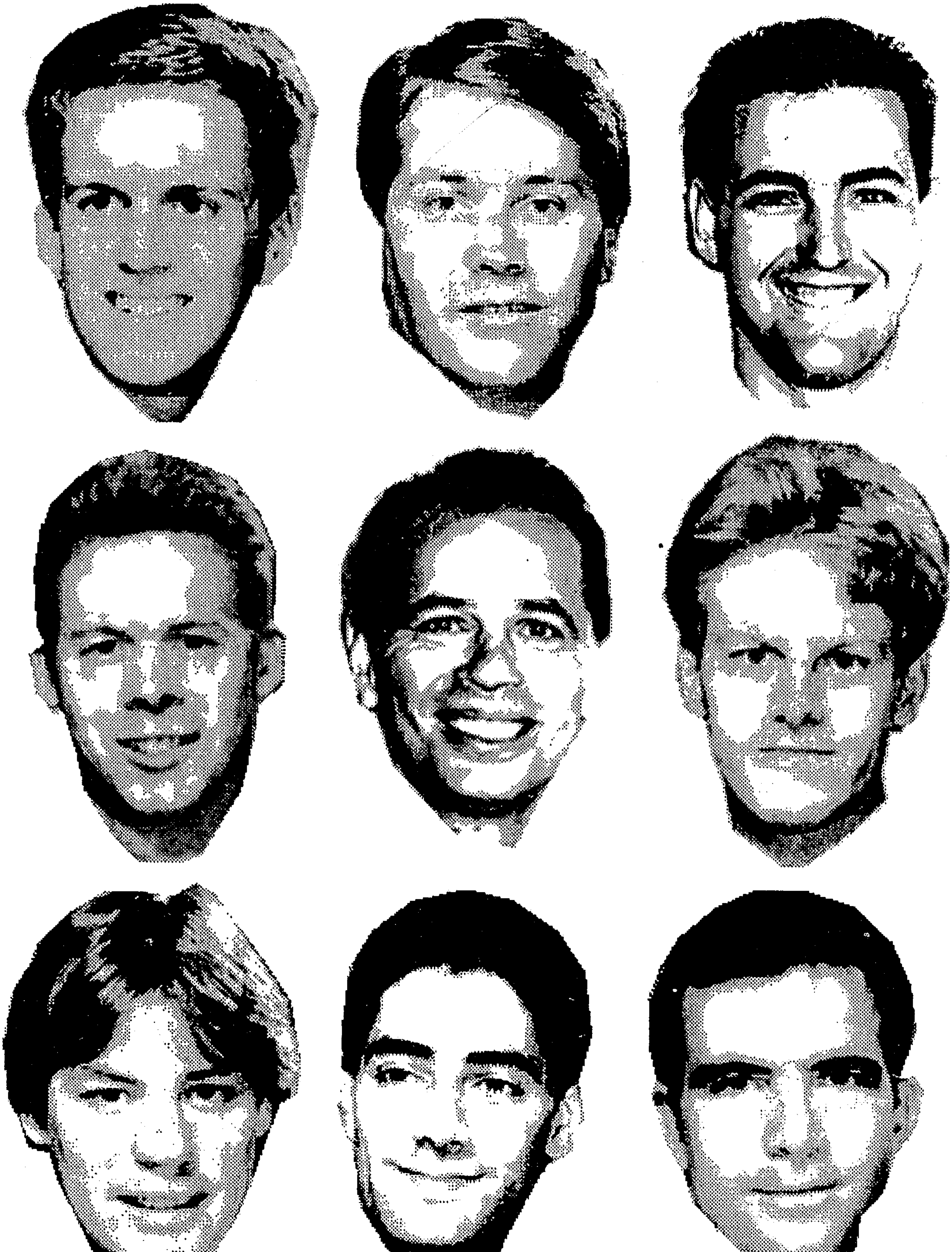


Figure 10D

Finally, in order to determine the quality of the composites, a set of 45 judges examined the final composites and then attempted to identify the culprit from the recognition set. Each judge evaluated no more than three different composites, and no judge evaluated more than one composite resulting from exposure to a particular culprit. Identification was assessed by requiring the judges to make 5 selections from the 36 faces in their order of preference. Judges were considered to have made an identification if the target face was included in their five selections.

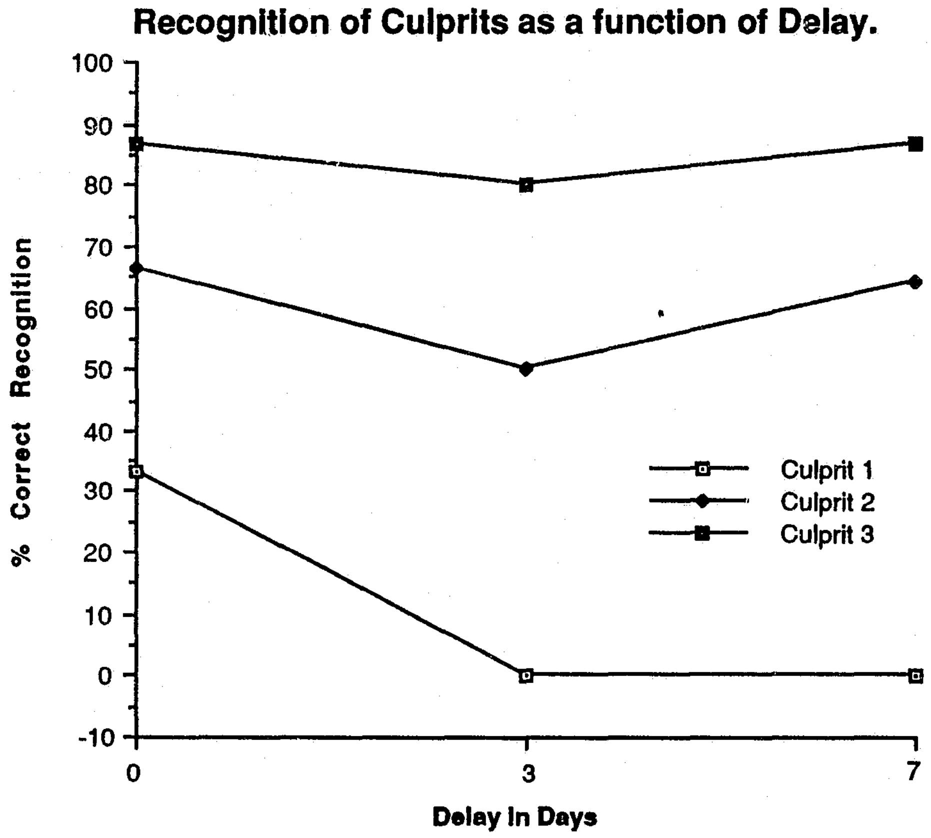


Figure 11: Recognition ability as a function of target culprit and delay.

Results: Figure 11 shows recognition performance across days, for the three culprits. (The Delay 0 data are the immediate recognition results from the pilot study.) An analysis of each delay group revealed a significant effect of target at Delay 0 ($X^2{}_2 = 9.26$; $p < .01$), Delay 3 ($X^2{}_2 = 16.86$; $p < .001$) and Delay 7 ($X^2{}_2 = 24.06$; $p < .001$). The only change in recognition performance over days was a decrement in the recognition of Culprit 1 ($X^2{}_1 = 10.9$; $p < .01$) over the three levels of delay. No such recognition decrement was observed for the other two culprits.

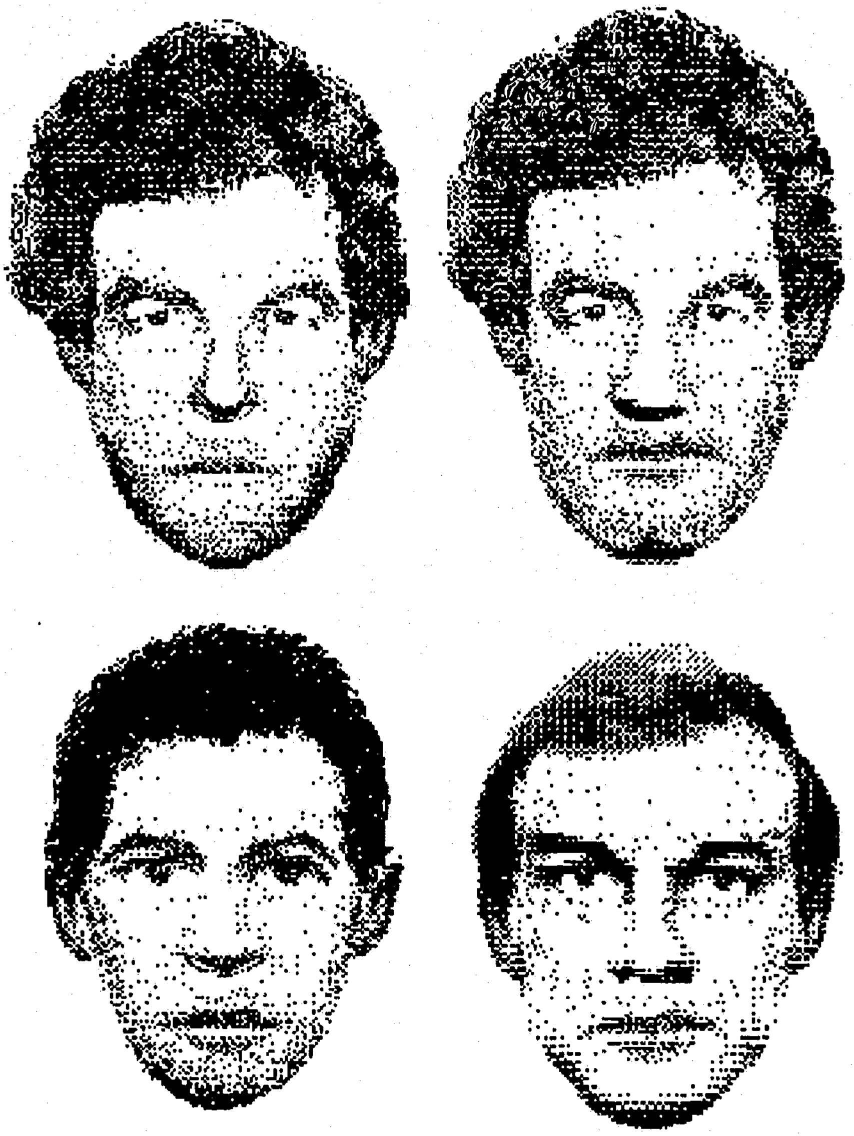


Figure 12: "Perfect composites" (left) and composites generated by witnesses (right) for culprit 2 (top) and culprit 1 (bottom). Both composites were generated by witnesses after a delay of 7 days from exposure to the simulated crime. Note that when the culprit has distinctive features (culprit 2), the composite is good; when a culprit has prototypical features (culprit 1), the composite is not useful.

The sex of the subject had no effect on recognition ability. In order to examine the relationship between the subject's recognition ability and subsequent identification of composites by the judges, the Delay 3 and Delay 7 subjects were divided into those with or without good recognition ability. This analysis revealed that subjects with good recognition produced composites which were correctly identified by judges at a significantly higher rates than chance ($X^2{}_1 = 18.98$; $p < .001$), and significantly higher rates than those with less recognition ability ($X^2{}_1 = 5.39$; $p < .05$). These subjects produced composites which led to identification by judges on 44% of their attempts. When Culprit 2 was the target, they produced composites which were identified on 45% of occasions; the identification rate was 43% when the target was Culprit 3. Since there was 0% recognition of Culprit 1 after a 3 or 7 day delay, this target generated poor composites which were never correctly identified by the judges. Figure 12 shows examples of composites generated when the target was either Culprit 1 or Culprit 2.

(G) DISCUSSION:

The FacePrints program uses a Genetic Algorithm to evolve a culprit's face by searching a space containing over 34 billion possibilities. The current version of the program uses bit-mapped graphics, a FreezeFeature option and cross-over and mutation rates which have been optimized for such a search. Simulated searches, using these features and parameters, have shown that an excellent composite can be evolved within ten generations if the simulated witness can (a) accurately recognize the culprit and (b) accurately rate a set of twenty faces according to their resemblance to the culprit.

It is useful to view the success or failure to generate a facial composite as being a consequence of a series of three major information losses, which occur between the time of witness exposure to the culprit and final identification of the culprit from a generated composite. The first loss can be attributed to witness recognition failure which may be a function of several variables; the conditions of the exposure, the distinctiveness of the target or the delay prior to generating the composite. A second loss occurs as a consequence of the process used for generating the composite. The process may lose information depending on whether it is based on recognition or recall, the completeness of the data base, the efficiency of the search or the adequacy of the user interface. A final loss occurs when a viewer attempts to identify the culprit on the basis of similarity to the generated composite. Even a "perfect" composite can fail to elicit identification (see pilot study 3). The task of recognizing the real culprit after having seen the composite, has much in common with the initial information loss experienced by the witness. As before, the nature of the exposure, the distinctiveness of the target and the delay between seeing the composite and seeing the target may all influence the recognition process. Since the purpose of the current study is to evaluate the FacePrints process (second loss) it is necessary to isolate these "process failures" from witness recognition failures (first loss) and failures associated with identification failures (third loss).

A major advantage of the FacePrints procedure over other methods for generating facial composites is its reliance on recognition rather than recall. If a witness is unable to recognize a culprit then accurate facial feature recall is not possible. However, a witness may recognize a culprit without possessing the ability to recall all, or even some, of the culprit's features. (We all recognize our mothers, or photographs of our mothers, but we may have difficulty describing her features). It is therefore important, when evaluating the FacePrints program, to assess the degree of each witness's recognition. If a witness is unable to recognize a culprit then the FacePrints process, (and almost certainly every other facial composite processes) will not adequately generate an accurate composite.

The current study indicates that culprit recognition is a function of the nature of the culprit and the time which has elapsed since a witness sees the culprit. Our results suggest that culprits with distinctive features are always recognized better than targets with features that are close to a prototypical face. Recognition of prototypical face is poor, even after having immediately viewed such a culprit, and under these circumstances recognition ability decays rapidly over time. In contrast, faces of non-prototypical culprits are well recognized, and there is no evidence for a decay in the recognition of such culprits over the seven day period examined in the current study. It should be noted that the recognition measurement employed in this experiment involved the recognition of a "natural-with-expression" photograph of the culprit resulting from a short videotaped exposure to the culprit's face. Thus, this decrement in performance can not be attributed to either the FacePrints process or the quality of the computer generated graphics. Process failures can only be determined by examining the success or failure to generate a composite in subjects who exhibit good recognition ability.

As noted above, the probability of generating an identifiable composite from a subject with good recognition ability is 0.44. (0.45 for Culprit 2 and 0.43 for Culprit 3). However, these probabilities include the identification losses, demonstrated in the third pilot study, as well as process failures. We can achieve a better estimate of the process effectiveness by removing such identification losses.

Let x = the probability of generating a good composite of a culprit using the Faceprints process. The probability of a poor composite = (1-x).

Even for a "perfect" composite, the probability of identifying this composite correctly from the recognition set varies with the nature of the culprit (Pilot Study 3). The p(identification) of Culprit 2 = 0.7; p(identification) of Culprit 3 = 0.70.

Therefore the probability of generating a Culprit 2 composite that is subsequently recognized is > ((0.70 * x) + the probability of a poor composite being identified correctly; (1-x)*5/36). Therefore:-

For Culprit 2 $0.70(x) + 0.14(1-x) >= 0.45$ (main experiment)
$x >= 0.55$
For Culprit 3 $0. 70 (x) + 0.14 (1-x) >= 0.43$ (main experiment)
$x >= 0.52$

Based on the data obtained from the current sample of subjects and culprits under laboratory conditions, the best estimate of the effectiveness of the FacePrints process is that it is capable of generating a useful composite of a criminal in more than 50% of cases, when the witness has good recognition ability.

(H) CONCLUSIONS:

FacePrints is the first recognition based method for generation facial composites. When a witness has good recognition ability then the current version of the program appears to be capable of generating a useful facial composite in more than 50% of cases. When a witness has poor recognition ability, then the FacePrints process (and probably every other facial composite system) is not a useful instrument.

Recognition ability appears to depend upon the distinctiveness of the culprit and the delay from exposure. Distinctive culprits are well recognized with little degradation over time (up to one week) whereas recognition of culprits with "prototypical" faces degrades quickly.

The sex of the witness does not appear to influence recognition ability or the effectiveness of the FacePrints process.

Since the composite generated by an eyewitness is often the only evidence available, it is important to develop methods which make maximum use of a witness's recognition ability. FacePrints is a step in this direction. This first series of studies have revealed several strengths and weaknesses of the process, which deserve more attention.

Strengths: The simulation studies have demonstrated that the GA is capable of searching a large "face space" (> 34 billion in the current experiments) and, with accurate feedback, can find a close resemblance to a culprit in a small number of generations. If necessary, this space could easily be expanded with little loss in efficiency. Enlarging the space, however, does not appear to be a critical variable since a small number of preliminary questions (e.g. the sex, color or approximate age of the culprit) can be used to specify which (34 billion) data base should be searched. A much more important factor is the construction of each data base (i.e which 512 hairs or 64 noses should it contain, and how they should be ordered). This problem is discussed below.

It has been demonstrated that the GA search can be (i) implemented using current computer technology (ii) used by untrained subjects and (iii) completed in less than one hour. All of these attributes are open to further improvement.

Weaknesses: An important weakness in FacePrints, and every other facial composite process, results from a lack of sufficient information concerning the relative values of the specific features and cephalometric measurements used in recognition. Haig (1984, 1986a) has provided some of this information, but much more is required in order to construct an adequate data base for practical use. When a feature is salient (e.g. chin shape, or nose-mouth distance), then a large number of alternatives must exist in the data base. Unimportant features (e.g. noses) require many fewer alternatives. Features also require organization along the relevant dimension(s) and separation by appropriate j.n.d. One approach to obtaining this data would be to expand the methodology used in the first pilot study (Figure 5). Determining the relative rates at which different features (and distances) are selected by subjects as they build a composite over generations, provides estimates of the relative importance that they place on these features. More systematic studies using different subjects and culprits could be used to generate the required information. Such datum would not only be important for generating better composites, it would also provide feature weights thus enabling the simulate witness program (SAM) to better represent the behavior of a real witness. Improvements to SAM would then permit more rapid progress in determining the best coding system and parameters for the GA search.

The current version of FacePrints uses bit-mapped images. Grey scale images would undoubtedly be better, but they would require much faster computing, more storage capacity, and present difficulties in image fusion when generating composites. An alternative would be to use an "air-diffusion" algorithm to compress grey scale images. Such "air-diffused" images maintain all of the quality of grey scaled images, without the undesirable characteristics noted above. Such "air-diffusion" algorithms are now readily available.

(I) <u>EXPECTED BENEFITS</u>:

Since the FacePrints program can be used like any other recall based process (e.g. Compusketch), it has all the advantages of other computerized methods. It has some unique advantages, however, which are listed below.

(1) Unlike all current facial recognition procedures, FacePrints depends upon recognition rather than recall. Since recognition is a highly developed skill in humans (Davies, Shepherd and Ellis, 1978), FacePrints should provide a more reliable means for generating composite pictures.

(2) The GA procedure is independent of the cognitive strategy employed by the witness. Both perceiver attributes, such as age, sex, hemispheric advantage and conceptual context, and processing strategies have been shown to influence facial recognition (Yarmey and Kent, 1980; Going and Reed, 1974; Hines, Jordan-Brown & Juzwin, 1987; Leehey, Carey, Diamond & Cahn, 1978; Goldstein and Chance, 1981; Wells, G.L. & Hryciw, 1984). However, since the GA does not enforce any particular strategy or rely upon any specific attribute, it allows a subject to pursue an individual approach. Because of the lack of constraints on the subject's processing method, the composites should be generated more efficiently and more accurately by the GA than by either the Compusketch method or from the assembly of composite parts. These advantages should be independent of the age, gender, hemispheric advantage or cognitive style of the subject.

(3) In addition to the selection and arrangement of facial elements, FacePrints provides a sensitive technique for examining the criteria which subjects use to generate and refine facial composites. Measuring the history of choices made by subjects as they search the multi-dimensional space of facial features, can provide specific information on the salient features and search strategies employed in facial recognition tasks. Such information is valuable for further refining the search procedure.

(4) The self-directed development of a suspect's face eliminates any biasing influences introduced through a human interview. Unlike other facial composite techniques, FacePrints does not require the use of an extensive set of questions about the suspect prior to generating the composite. This reduces the possibility that other information, unintentionally provided by the questioner, may bias witnesses in their selection of facial features. Because of this, the GA approach promises to be a more valid instrument.

(5) Interactions between features and their positions may be a major source of error when features are selected and then position-adjusted in two separate operations. This is the common strategy in current computerized systems. By representing facial variables as genes, both cephalometric measurements and specific feature elements of the composite may be coded in the same genotype. Witnesses using the GA can therefore evolve both the facial features and their relative positions at the same time, and in context.

(6) The use of a common gene code allows additional attributes, such as color, to be added easily in any future development of the FacePrints process.

(7) When used to implement a selection routine for facial identification, FacePrints provides a selection strategy that performs the double function of generating a composite and a genotype for that composite. This genotype can serve as a code for that individual face, not

unlike a fingerprint. These genotypes are potentially useful codes for comparing composite faces with stored records.

(8) Genotypes generated by a number of witnesses may be combined and used to generate a new composite face. These may prove to be more reliable than single source composites.

(9) FacePrints, as currently implemented, is designed so that minimal training is required by either the law enforcement officer or the witness. No artistic ability or computer expertise in necessary to generate a composite.

(10) The final version of FacePrints is expected to run on a Macintosh SE/30 computer. This should provide a facial composite process within the budget of most local law enforcement agencies.

(J) REFERENCES:

Baker, J.E. (1987). Reducing bias and inefficiency in the selection algorithm. Genetic algorithms and their applications: Proceedings of the second international conference on genetic algorithms, 'p 14-21.

Bartlett, F.C. (1932). Remembering. Cambridge, England: Cambridge University Press.

Brewer, W.F., & Nakamura, G.V. (1984). The nature and function of schemas. In R.S. Wyer, Jr., & T.K. Srull (Eds.), Handbook of Social Cognition: Vol 1 (pp. 119 - 160). Hillsdale, NJ: Lawrence Ehlbaum.

Cook, M. (1978). Eyemovements during recognition of faces. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.) Practical aspects of memory. (pp. 287-292). London: Academic Press.

Davidoff, J.B. (1986). The specificity of face perception: Evidence from psychological investigations. In R. Bruyer (Ed.), The neuropsychology of face perception and facial expression, (pp. 147-166). New Jersey: Lawrence Eribaum Associates.

Davies, G.M. & Christie, D. (1982). Face recall: An examination of some factors limiting composite production accuracy. Journal of Applied Psychology, 67, 103-109.

Davies, G.M., Ellis, H.D. & Shepherd, J.W. (1978). Face identification - the influence of delay upon accuracy of Photo-Fit construction. Journal of Police Science and Administration, 6(1), 35-42.

Davies, G.M., Shepherd, J.W. & Ellis, H.D. (1978). Remembering faces - acknowledging our limitations. Journal of the Forensic Science Society, 18, 19-24.

Dawkins, R. (1986). The blind watchmaker. New York: W.W. Norton & Company.

Ellis, H.D., Davies, G.M. & Shepherd, J.W. (1986). Introduction: Processes underlying face recognition. In R. Bruyer (Ed.), The Neuropsychology of face perception and facial expression (pp.1-38). New Jersey: Lawrence Erlbaum Associates.

Freeman, J.M., & Ellis, H.D. (1984). The effects of stimulus and subject factors on a face matching task. Neuropsychologia, 22, 635-638.

Going, M. & Read, J.D. (1974). Effects of uniqueness, sex of subject, and sex of photograph on facial recognition. Perceptual & Motor Skills, 39(10), 109-110.

Goldberg, D.E. (1989). Genetic algorithms in search, optimization & machine learning. Massachusetts: Addison-Wesley.

Goldstein, A.G., & Chance, J.E. (1981). Laboratory studies of face recognition. In G. Davies, H. Ellis, & J. Shepherd (Eds.), Perceiving and Remembering Faces. (pp. 81-104). New York: Academic Press.

Hagen, M.A. & Perkins, D. (1983). A refutation of the hypothesis of the superfidelity of caricatures relative to photographs. Perception, 12, 55-61.

Haig, N.D. (1984). The effect of feature displacement on face recognition. Perception, 13, 505-512.

Haig, N.D. (1986a). Exploring recognition with interchanged facial features. Perception, 15, 235-247.

Haig, N.D. (1986b). High-resolution facial feature saliency mapping. Perception, 15, 373-386.

Hall, D.F. (1976). Obtaining eyewitness identification in criminal investigations - applications of social and experimental psychology. Dissertation: Ohio State University.

Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor: University of Michigan Press.

Hines, D., Jordan-Brown, L. & Juzwin, K.R. (1987). Hemispheric visual processing in face recognition. Brain & Cognition, 6(1), 91-100.

Kennedy, D.F., Scannapieco, C.C., Mills, S.M. & Carr, W.J. (1985). Differential recognition of the right vs. left halves of human faces. Bulletin of the Psychonometric Society, 23(3), 209-210.

Kitson, T., Darnbrough, M. & Shields, E. (1978). Let's face it. Police Research Bulletin, Spring (30), 7-13.

Laughery, K.R., Fowler, R.H., & Rhodes, B.T. (1976). Mug file project report number UHMUG-3 - Factors affecting facial recognition. (Grant no. 76-NI-99-012). Rockville, MD: National Institute of Justice / National Criminal Justice Reference Service Microfiche Program.

Leehey, S., Carey, S., Diamond, R. & Cahn, A. (1978). Upright and inverted faces: The right hemisphere knows the difference. Cortex, 14(3), 411-419.

Light, L.L., Kayra-Stewart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. Journal of Experimental Psychology: Human Learning and Memory, 5, 212-218.

Loftus, E.F. & Greene, E. (1980). Warning - even memory for faces can be contagious. Law and Human Behavior, 4, 323-334.

Mason, S.E. (1986). Age and gender as factors in facial recognition and identification. Experimental Aging Research, 12(3), 151-154.

McClelland, J.L., & Rumelhart, D.E. (1965). Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114, 159-188.

Miller, L.K. & Barg, M.D. (1983). Dissociation of feature vs. configural properties in the discrimination of faces. Bulletin of the Psychonomic Society, 21(6), 453-455.

Neumann, P.G. (1974). An attribute frequency model for the abstraction of prototypes. Memory and Cognition, 2, 241-248.

Neumann, P.G. (1977). Visual prototype formation with discontinuous representation of dimensions of variability. Memory and Cognition, 5, 187-197.

Penry, J. (1974). Photo-Fit. Forensic Photography, 3(7), 4-10.

Posner, M.I. (1973). Cognition: An Introduction. Glenview, Illinois: Scott, Foresman.

Rhodes, G. (1985). Perceptual asymmetries in face recognition. Brain & Cognition, 4(2), 197-218.

Rhodes, G., Brennan, S. & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representation of faces. Cognitive Psychology, 19, 473-497.

Ross-Kossak, P. & Turkewitz, G. (1986). A micro and macro developmental view of the nature of changes in complex information processing: A consideration of changes in hemispheric advantage during familiarization. In R. Bruyer (Ed.), The neuropsychology of face perception and facial expression, (125-145). New Jersey: Lawrence Erlbaum Associates.

Solso, R.L. & McCarthy, J.E. (1981). Prototype formation of faces: A case of pseudo-memory. British Journal of Psychology, 72, 499-503.

Valentine, T. & Bruce, V. (1986). The effects of distinctiveness in recognizing and classifying faces. Perception, 15, 525-535.

Wells, G.L. (1978). Applied eyewitness testimony research: System variables and estimator variables. Journal of Personality and Social Psychology, 36(12), 1146-1557.

Wells, G.L. & Hryciw (1984). Memory for faces: Encoding and retrieval operations. Memory and Cognition, 12 (4), 338 - 344.

Wells, G.L., & Wright, E.F. (1986). Practical issues in eyewitness research. In M.F. Kaplan (Ed.), The impact of social psychology on procedural justice. Springfield, IL: C.C. Thomas.

Yarmey, A.D. & Kent, J. (1980). Eyewitness identification by elderly and young adults. Law and Human Behavior, 4(3), special issue, 359-371.