

EvoFIT: A Holistic, Evolutionary Facial Imaging System

Charlie David Frowd

Submitted for the degree of Doctor of Philosophy, Department of Psychology, University of Stirling, April 2001.

Acknowledgements

There are many people who have been of great assistance to me during my Ph.D. Firstly, I should like to thank the EPSRC for funding this project¹ and my supervisor, Peter Hancock, for his constant support and patience with my crazy ideas and enthusiasm. Along with Vicki Bruce, I'd like to thank the many people of the Faces Lab for their comments, suggestion and input: Zoë Henderson, Karen Lander, Chris O'Donnell, Steve Langton, Helen Honeyman, Craig Newman, Jenny Rarity, Hayley Ness and Fraser Smith. I am also very grateful to Derek Carson for the collaborative work that we have done. Next, I should like to thank the University's CCCN members for asking me to speak on several occasions and feedback they have provided. I'd like to make a special mention also to Andy Rolph of Grampian Police who has shown continued support and enthusiasm for the project from the onset. I am also grateful to him for inviting me to the ACPO(S) meetings, the result of which has provided invaluable forensic input. I am also very grateful to Paul Spick of Northampton Police for the opportunity to apply the system to a real case. My thanks also go to those that have worked on the project over the years, including Clare, Jamie, Carole, Kathy, Kirsty, Kirsty, Jason, Kerry, Kim, Kristine, Sarah and David. Last, but by no means least, I should like to thank my very good friend Helen Honeyman and my family, all of whom have tolerated so many of my ramblings about face perception and the EvoFIT "way of thinking".

¹ EPSRC grant number GR/L88627.

Contents

LIST OF FIGURES	9
LIST OF TABLES	12
LIST OF EXPERIMENTS & SIMULATIONS	13
ABSTRACT	14
CHAPTER 1: REVIEW AND APPROACH	15
Review of Photofitting	15
The Manual Systems	15
The Electronic Systems	19
Summary	23
Holistic Notions	24
Parametrized Models	27
Which Method(s) to Adopt?	31
Towards a New Approach	31
Conclusion	35
CHAPTER 2: PILOT WORK (MARK I FACE EVOLVER)	36
Design Considerations	36
Computer System	36
Software	37
Pilot Design	40
Constructing a Holistic Face Model	42
Experiment 1: Searching the Model	44
Method	45
Participants	45
Apparatus	45
Targets	45

Procedure	46
Results	47
Mean Square Error (MSE)	47
General Analysis	54
Rating	56
Timing data	57
Discussion	57
Summary and Further Work	60
CHAPTER 3: FULL-FACE SYSTEM (MARK II FACE EVOLVER)	63
Increasing Utility	63
Design of a Mark II System	64
Face Shape Model	64
Hair	64
Targets	65
Improving Image Quality	65
Demo System	68
Experiment 2: Shape-Variant Face Model	74
Method	74
Participants	75
Apparatus	75
Procedure	75
Results	76
Discussion	82
Experiment 3: Confirmation of Rating Scores	84
Method	84
Participants	84
Procedure	84
Results	85
Discussion	85
Experiment 4: Celebrity Targets	86
Method	86
Participants	88
Apparatus	88
Procedure	88
Results	89

Discussion	91
Experiment 5: Recognition of Evolved Celebrities	92
Method	92
Participants	93
Procedure	94
Results	94
Discussion	99
Experiment 6: Appropriacy of the Hairstyle	99
Method	99
Participants	100
Procedure	100
Results	100
Discussion	101
Experiment 7: Increasing the Population Size, Random Targets	102
Method	102
Participants	102
Results	102
Experiment 8: Increasing the Population Size, Celebrity Targets	103
Method	103
Participants	103
Results	103
Comparison of Chapter Experiments	104
Discussion	105
General Discussion	106
CHAPTER 4: SIMULATIONS	110
Appropriacy of Parameter Settings	110
The Simulations	111
Simulation 1: Population Size	111
Simulation 2: Selection Pressure	115
Simulation 3: Mutation Rate	116
Simulation 4: Number of Selected Faces	118

Simulation 5: Elitism	120
Simulation 6: Combined Effects with Elitism Enabled	122
Simulation 7: Population Size (Revisited)	122
Simulation 8: Coefficient Pruning	124
Simulation 9: Separate Selection Mechanisms	126
General Discussion	129
Parameter settings	130
CHAPTER 5: THE EVOFIT SYSTEM	132
What Must Be Done?	132
Multiple Face Palettes	132
Experiment 9: Separate Shape and Texture Selection	136
Results	141
Discussion	141
More Palettes?	141
Increasing the Complexity of the Face Model	145
The Problem of Unwanted Change in Pose	146
The Feature Shifter	147
Hair and Overlays	149
Summary of Developments to the Mark II Face Evolver	152
Experiment 10: Evaluating the EvoFIT System	153
Method	153
Selection of Target Stimuli	154
Creating the Computer-Generated Composites	155
Participants	156
Apparatus	156
Procedure	157
Recognizing the Composites	157
EvoFITs	158
Participants	158

Results	158
EFITs	159
Participants	159
Results	159
Comparison with EvoFIT	161
Discussion	161
Experiment 11: EvoFITs of Young Famous Faces	163
Creation of the EvoFITs	163
Evaluation of the EvoFITs	163
Results	164
Discussion	164
Operation Mallard	164
Verifying Anonymity	168
General Discussion	170
Conclusion	171
CHAPTER 6: FUTURE WORK	172
Summary of Previous Work	172
Evolution from Memory	173
Face Model Issues	174
Representation	174
Colour	174
¾ View	175
Additional Databases	176
Anonymity	176
Non-Holistic Bias	177
Verbal Description	177
Feature Shifter	178
Anchored Face Similarity Scale	179
Holistic Theory	180
Enhancing Performance Still Further	180
Simulations	180
Feature Shifter	181
Eyes and Hair	182

Multiple Witnesses	183
Final Comments	184
GLOSSARY OF ABBREVIATIONS	185
REFERENCES	186
APPENDICES	200
Appendix A: Mean Square Error Measures of Facial Images	201
Appendix B: Famous Face Stimuli Used for Experiment 10	203
Appendix C: EFIT Description Sheet Used for Experiment 10	204
Appendix D: Composites Created in Experiment 10	206
(a) Composites Created by EFIT	206
(b) Composites Created by EvoFIT	214
Appendix E: Targets used for Experiment 11	222
Appendix F: EvoFIT Operating Procedures for Experiment 11	223
Appendix G: EvoFITs created in Experiment 11	242
Appendix H: Flowcharts for Face Generation and EvoFIT Operation	245

List of Figures

<i>Figure 1: An Example Alignment of Facial Feature Coordinate Points</i>	43
<i>Figure 2: Examples of Shape-Normalized Faces used to Construct Face Model</i>	43
<i>Figure 3: The Target Faces</i>	46
<i>Figure 4: Performance of Subject 3, Target 1</i>	47
<i>Figure 5: Performance of Subject 17, Target 1</i>	48
<i>Figure 6: Overall Mean Subject Performance (between First and Last Generation)</i>	49
<i>Figure 7: Correlation between MSE and Ratings (all Targets)</i>	49
<i>Figure 8: Average Rating Scores with the Corresponding Reduction in Average MSE for Each Subject (the graph is ordered by increasing rating)</i>	50
<i>Figure 9: No Trend Between Mean Rating and Mean Reduction in MSE</i>	51
<i>Figure 10: Positive Trend between the Standard Deviation of Rating Scores and the Average Reduction in MSE</i>	52
<i>Figure 11: Subject 14's Rating-MSE Correlation and Reduction in MSE for Target 2</i>	53
<i>Figure 12: Cumulative Reduction in Average MSE (all Targets)</i>	54
<i>Figure 13: Mean MSE for all Targets (graph bars indicate SD of MSE)</i>	55
<i>Figure 14: Distribution of Maximum Ratings Assigned For Each Target</i>	56
<i>Figure 15: Average MSE Score for Faces Rated as Maximum</i>	56
<i>Figure 16: Average Time Taken to Rate a Population of Faces</i>	57
<i>Figure 17: Coordinate Point Locations for the Full Shape-Variant Model</i>	66
<i>Figure 18: Keying mask used to create a composite image from the internal and external features</i>	67
<i>Figure 19: Examples of Randomly Generated Shape-Variant Faces</i>	68
<i>Figure 20: Presentation Format for Faces (the target is displayed in the centre of the population)</i>	70
<i>Figure 21: User Message Presented at Start of Experiment</i>	75
<i>Figure 22: User Message Presented In Front of the First Set of Faces</i>	76
<i>Figure 23: User Message Presented In Front of the Second Set of Faces</i>	76
<i>Figure 24: Examples of Evolutionary Performance</i>	77
<i>Figure 25: Improvement in Rating Scores</i>	78
<i>Figure 26: Distribution of the highest rating for each participant</i>	79
<i>Figure 27: Distribution of the Highest Rated Faces</i>	80
<i>Figure 28: Proportion of Time that Participant Ratings Increased</i>	80
<i>Figure 29: Increase in Peak Rating Scores Over Increasing Longer Generations</i>	81
<i>Figure 30: Variability in Mean and Standard Deviation of MSE during Evolution</i>	82
<i>Figure 31: Famous Male Celebrities used as Targets. These are (left to right, top to bottom): Robbie Williams, Tim Henman, Pierce Brosnan, Robert Carlyle, George Clooney and Hugh Grant</i>	87
<i>Figure 32: The First Six Hairstyles Available for Selection</i>	88
<i>Figure 33: Mean Rating Scores for the 6 Celebrities (data from Subject 5 omitted)</i>	89
<i>Figure 34: Distribution of the Highest (Peak) Rating for Each Participant</i>	90
<i>Figure 35: Decrease in MSE Scores with Increasing Generation</i>	91

<i>Figure 36: Distribution of Correct Guesses in the Forced-Choice Identification Task</i>	95
<i>Figure 37: Examples of the Most Frequently Recognized Evolved Targets (identification at least 2/8 or 25%; ratings assigned during evolution)</i>	96
<i>Figure 38: Unsuccessfully Identified Evolved Targets</i>	98
<i>Figure 39: Hairstyle Ratings for each Evolved Celebrity (Group A in white bars, Group B in grey bars)</i>	101
<i>Figure 40: Additional Celebrities used For Experiment 8. These are (left to right, top to bottom): for Brad Pitt, Bryan Adams, John Travolta, Timothy Dalton, Alec Baldwin and Ewan McGregor</i>	103
<i>Figure 41: Comparison of Rating Scores between Experiment 2, Experiment 4, Experiment 7 and Experiment 8</i>	104
<i>Figure 42: Rating Performance for Target Type and Population Size (data from generations 1 to 4 only)</i>	105
<i>Figure 43: Effect of Varying Population Size on Average MSE (1:1 selection pressure, mutation probability of 0.0, no elitism and $\sqrt{6/16}$*population size]selected faces)</i>	114
<i>Figure 44: Effect of Varying Population Size on Average Minimum MSE (1:1 selection pressure, mutation probability of 0.0, no elitism and $\sqrt{6/16}$*population size]selected faces)</i>	115
<i>Figure 45: Effect of Varying Selection Pressure (of best face) on Average MSE (16 population faces with 6 faces selected per generation, mutation probability of 0.0 and no elitism)</i>	116
<i>Figure 46: Effect of Varying Selection Pressure (of best face) on Average Minimum MSE (16 population faces with 6 faces selected per generation, mutation probability of 0.0 and no elitism)</i>	116
<i>Figure 47: Effect of Varying Mutation on Average MSE (16 population faces with 6 faces selected per generation, no elitism and 2:1 selection pressure)</i>	117
<i>Figure 48: Effect of Varying Mutation on Average Minimum MSE (16 population faces with 6 faces selected per generation, no elitism and 2:1 selection pressure)</i>	118
<i>Figure 49: Effect of Varying the Number of Faces Selected on Average MSE (16 population faces, mutation probability of 0.1, no elitism and 2:1 selection pressure)</i>	119
<i>Figure 50: Effect of Varying the Number of Faces Selected on Average Minimum MSE (16 population faces, mutation probability of 0.1, no elitism and 2:1 selection pressure)</i>	119
<i>Figure 51: Initial Effect of Enabling Elitism on Average Minimum MSE (16 population faces, mutation 0.1 and 2:1 selection pressure)</i>	120
<i>Figure 52: Effect of Repaired Elitism Mechanism on Average Minimum MSE (16 population faces, mutation 0.1 and 2:1 selection pressure)</i>	121
<i>Figure 53: Effect of the Elitism Mechanism on Average MSE (16 population faces, mutation 0.1 and 2:1 selection pressure)</i>	121
<i>Figure 54: Effect of Varying the Population Size on Average Minimum MSE (mutation 0.1, 2:1 selection pressure, 6 faces selected and elitism enabled); the white bars indicate a non-significant difference compared with a population size of 16 faces.</i>	124
<i>Figure 55: Effect of Shape Pruning on Average Minimum MSE (16 population faces, mutation 0.1, elitism enabled and 2:1 selection pressure)</i>	126
<i>Figure 56: Effect of Texture Pruning on Average Minimum MSE (16 population faces, mutation 0.1, elitism enabled and 2:1 selection pressure)</i>	126

<i>Figure 57: Effect of the Maximum Mutation Rate (probability of 1.0) on Average Minimum MSE (16 population faces, mutation 0.1, elitism off and 2:1 selection pressure)</i>	127
<i>Figure 58: Effect of Separate Evolution for Shape and Texture Components on Average MSE (16 population faces, mutation 0.1, elitism off and 2:1 selection pressure)</i>	128
<i>Figure 59: Effect of Separate Evolution for Shape and Texture Components on Average Minimum MSE (16 population faces, mutation 0.1, elitism off and 2:1 selection pressure)</i>	129
<i>Figure 60: An Example Facial Shape Palette (FSP) Displaying a set of Randomly Generated Shapes.</i>	134
<i>Figure 61: An Example Facial Texture Palette (FTP) Displaying a set of Randomly Generated Textures</i>	135
<i>Figure 62: EvoFITs Created with Target Visible (A), from Inspection (B) and from Memory (C)</i>	138
<i>Figure 63: An Example Facial Shape Palette with Best Texture (FSPBT) for a set of Randomly Generated Shapes</i>	143
<i>Figure 64: An Example Facial Texture Palette with Best Shape (FTPBS) for a set of Randomly Generated Textures</i>	144
<i>Figure 65: An Example of an Unwanted Pose Change</i>	146
<i>Figure 66: Removing the Unwanted Pose Change</i>	147
<i>Figure 67: An Example Illustrating the Effect of Moving the Eyes Closer into Register (by 8 pixels). The original image is on the left</i>	148
<i>Figure 68: Examples of Hairstyles Imported from the PROfit Package and Applied to a Population Face</i>	151
<i>Figure 69: An Example of Adornments Added to the Internal and External Facial Features</i>	152
<i>Figure 70: Conditional Hit Rate (CHR) of EvoFITs Grouped by Distinctiveness (names are sorted by surname within each category)</i>	158
<i>Figure 71: Conditional Hits Rate for Low, Medium and High Distinctive EvoFITs</i>	159
<i>Figure 72: Conditional Hit Rate of EFITs Grouped by Distinctiveness (names are sorted by surname within each category)</i>	160
<i>Figure 73: Conditional Hits Rate for Low, Medium and Highly Distinctive EFITs</i>	160
<i>Figure 74: CHR of EvoFITs Ordered by Age of Famous Person</i>	162
<i>Figure 75: Original Artist's Sketch</i>	165
<i>Figure 76: Updated Sketch of Hair</i>	165
<i>Figure 77: The Updated Sketch Used as the External Facial Features</i>	166
<i>Figure 78: The EvoFIT Constructed in the Field Test</i>	166
<i>Figure 79: Comparison of the Spatial Relationship between the Original Sketch and the EvoFIT: (a) Alignment with the Eyes (b) Alignment with the Mouth. The outline of the EvoFIT features is superimposed on the artist's sketch</i>	168
<i>Figure 80: RMS Error for Shape between the EvoFIT and Database Images</i>	169
<i>Figure 81: RMS Error for Shape-Free Texture of the Internal Features between the EvoFIT and Database Images</i>	169
<i>Figure 82: Imperfections in the Generated Images: distortions to the Irises caused by Morphing and the Problem of "Floating" hair</i>	182
<i>Figure 83: Distribution of MSE Scores for the Pilot System</i>	201

List of Tables

<i>Table 1: Performance of Subject 14 on Target 2, Generation 3</i>	<i>53</i>
<i>Table 2: The Anchored Face Similarity Scale (AFSS) used to Evaluate Performance</i>	<i>73</i>
<i>Table 3: Distribution of Celebrities Chosen for Evolution.....</i>	<i>89</i>
<i>Table 4: List of Celebrities Used in the Forced-Choice Task (an asterisk * indicates celebrities that were targets in Experiment 4, the remaining items are foils)</i>	<i>93</i>
<i>Table 5: The Anchored Hairstyle Similarity Scale</i>	<i>100</i>
<i>Table 6: Summary of Suggested Parameter Values (values in bold indicate new settings and features).....</i>	<i>131</i>

List of Experiments & Simulations

<i>Experiment 1: Searching the Model</i>	44
<i>Experiment 2: Shape-Variant Face Model</i>	74
<i>Experiment 3: Confirmation of Rating Scores</i>	84
<i>Experiment 4: Celebrity Targets</i>	86
<i>Experiment 5: Recognition of Evolved Celebrities</i>	92
<i>Experiment 6: Appropriacy of the Hairstyle</i>	99
<i>Experiment 7: Increasing the Population Size, Random Targets</i>	102
<i>Experiment 8: Increasing the Population Size, Celebrity Targets</i>	103
<i>Experiment 9: Separate Shape and Texture Selection</i>	136
<i>Experiment 10: Evaluating the EvoFIT System</i>	153
<i>Experiment 11: EvoFITs of Young Famous Faces</i>	163
<i>Simulation 1: Population Size</i>	111
<i>Simulation 2: Selection Pressure</i>	115
<i>Simulation 3: Mutation Rate</i>	116
<i>Simulation 4: No of Selected Faces</i>	118
<i>Simulation 5: Elitism</i>	120
<i>Simulation 6: Combined Effects with Elitism Enabled</i>	122
<i>Simulation 7: Population Size (Revisited)</i>	122
<i>Simulation 8: Coefficient Pruning</i>	124
<i>Simulation 9: Separate Selection Mechanisms</i>	126

Abstract

This thesis details the development and evaluation of a new photofitting approach. The motivation for this work is that current photofit systems used by the police – whether manual or computerized – do not appear to work very well. Part of the problem with these approaches is they involve a single facial representation that necessitates a verbal interaction. When a multiple presentation is considered, our innate ability to recognize faces is capitalized (and the potentially disruptive effect of the verbal component is reduced). The approach works by employing Genetic Algorithms to evolve a small group of faces to be more like a desired target. The main evolutionary influence is via user input that specifies the similarity of the presented images with the target under construction.

The thesis follows three main phases of development. The first involves a simple system modelling the internal components of a face (eyes, eyebrows, nose and mouth) containing features in a fixed relationship with each other. The second phase applies external facial features (hair and ears) along with an appropriate head shape and changes in the relationship between features. That the underlying model is based on Principal Components Analysis captures the statistics of how faces vary in terms of shading, shape and the relationship between features. Modelling was carried out in this way to create more realistic looking photofits and to guard against implausible featural relationships possible with traditional approaches. The encouraging results of these two sections prompted the development of a full photofit system: EvoFIT. This software is shown to have continued promise both in the lab and in a real case. Future work is directed particularly at resolving issues concerning the anonymity of the database faces and the creation of photofits from the subject's memory of a target.

(292 words)

Chapter 1: Review and Approach

This chapter reviews the wealth of research investigating the utility of photofit systems. The manual systems are examined first and it is found that Identikit and Photofit are inherently limited by the available facial features. The computerized versions (e.g. Mac-a-mug and EFIT) have greater expression, but this advantage diminishes when attempts are made to create composites from memory (as necessary in use). It is shown that the feature-based approach adopted in all these systems is not only at odds with the holistic way faces are perceived generally but with strategies that are frequently adopted to remember a face. A novel approach is proposed that does not inherently decompose a face into its component parts. The approach creates a photofit by making inherently non-verbal judgments based on the presentation of multiple faces seen at the same time. Genetic Algorithms are used to “breed” these choices until an acceptable likeness is reached. An additional technique is suggested for situations requiring a more feature-based method of construction.

Review of Photofitting

Traditionally, a photofit is a visual representation of an assailant composed from a set of predefined facial parts. In a forensic setting, a verbal description of an assailant would be obtained from a witness to a crime. A photofit “operator” would then select the most likely combination of facial features that match the verbal description. This so-called facial “composite” would then be presented to the witness. The witness would then suggest changes necessary to this face until an acceptable likeness has been reached (Davies, Shepherd, Shepherd, Flin & Ellis, 1986).

There are two broad systems for generating composites: the manual approaches, where a face is assembled by hand, and the computerization thereof.

The Manual Systems

Of the major manual systems, there are two well-known types: Identikit and Photofit². In the Identikit system, facial features are printed on acetate transparencies. There are five sets

² There is considerable variation in the literature regarding the spelling of photofit products. For example, the “Photofit” system has been written as Photofit, Photo-fit, Photo-Fit and even Photo-FIT. The style used in Shepherd & Ellis’s (1996) excellent review of photofitting systems has been adopted throughout this thesis. Hence, the terms Photofit, Identikit, EFIT and Mac-a-mug refer to specific photofit systems. Note that when I write “photofit” (i.e. with a lower case letter ‘p’), this refers to a photofit created from any photofit system (including EvoFIT). The only other generic term used in this thesis for “photofit” is “composite”. Note that the latter term has quite a general meaning that also includes representations

of facial features available: forehead and hair; eyes and eyebrows; nose; mouth and lips; and chins. A composite is constructed by laying acetates on top of each other. These “slides” can be exchanged until an acceptable likeness is obtained. Shepherd & Ellis (1996) explain that when first released by P.J. Dunleavy in 1959, features comprised of line drawings, but a more recent version in 1975 (called Identikit II) contained photographic elements. Whereas the Identikit system was used primarily in the US, Photofit was primarily adopted in this country.

The Photofit system is similar to Identikit II in that photographic elements are used. However, rather than acetates, facial features in Photofit are printed on jigsaw-like pieces that slot into a template. The system was released in the late 60s and early 70s (Penry, 1970; and Penry, 1974). There are 550 facial features in this system, roughly the same as the 470 available in Identikit II (Shepherd & Ellis, 1996). Both systems make use of a marking pencil to elaborate facial features.

Much research has been carried out in the 1970s and 1980s to establish the effectiveness of these systems. In general, this research has not been very positive. In one of the first studies, Ellis, Shepherd & Davies (1975) had subjects build composites using the Photofit system with the target present during construction or following a 10 second exposure. The resulting composites were rated for likeness to their target by independent judges on a 7-point scale. Rating scores were significantly higher when the target was visible during construction compared with photofits constructed from memory. In a following experiment, 12 of the original 32 subjects - the six who had performed the best and the six who had performed the worst - made new composites from memory. A different group of participants attempted to pick out the original photographs from 35 distractors with the composites on display. Overall performance was rather low at 12.5% correct, although composites made from subjects who were found to be good encoders in the previous experiment [those with higher rated composites] had a significantly higher success rate (16.2%) than the poor encoders.

A low level of success was also observed by Ellis, Davies & Shepherd (1978a). Subjects constructed composites with differing target exposure (15 seconds or 2.5 minutes) and either intentional or non-intentional face learning. The resulting composites were rated on a 7 point scale by a different group of subjects and no significant differences were found across conditions. Overall, the average rating³ scores corresponded to the category between “below average” and “moderate” likeness to the target and were thus lower than the available categories of “good” and “very good”. Further problems with the Photofit system were

produces from a Sketch Artist: “the term composite means any image produced by computerised systems or drawn” (ACPO(S), 2000, page 14).

³ Throughout this thesis, unless otherwise specified, the average of a variable is assumed to refer to the *mean*. Frequently, angle brackets are used to represent the mean of a variable. For example, <rating> would refer to mean rating scores.

illustrated by no significant differences in rating scores of photofits created with the target visible and from memory. It is believed that the ability to constantly scrutinize a face [i.e. in the target visible condition] should enable finer adjustments to be made and result in an overall better rated composite. This was not found to be the case and the authors imply that it is the system that is limiting performance rather than the subject's memory of a target.

A similar discovery has been made using the Identikit system⁴. Laughery & Fowler (1980) observed no significant difference in rating scores between Identikit composites constructed with the target present and from memory. Their study employed the use of a sketch artist - often used as an alternative representation to a composite in a forensic setting - to create a drawing for each of their 71 target faces. They had reason to believe that the targets were not too difficult to represent visually since the sketches were rated significantly better [constructed with the target present or from memory] than the composites.

Note that, like the photofit systems, the sketch artist typically works from a description provided by a witness (Shepherd & Ellis, 1996). The effectiveness of the verbal description itself has been compared against the manual systems. In Christie & Ellis (1981) for example, subjects viewed a target for 60 seconds, provided a description and then constructed a composite with the Photofit system. A different group of subjects attempted to identify all six targets in 18 distractors given the description or the composite. The results clearly showed that verbal descriptions were superior (48% accuracy in this matching task) to photofits (23% accuracy).

It is worth noting at this stage that the utility of a Photofit is believed to be primarily to limit the number of possible suspects, a so-called "type likeness" (e.g. Ellis, 1996; and Penry, 1976), rather than to actually identify them. This notion is reflected in a study that asked police officers how valuable they thought Photofit was in solving 140 crimes (Darnborough, 1977 [cited in Clifford & Davies, 1989]). In only 5% of cases the photofit was "entirely responsible" and in 17% was the photofit "very useful". A later study in 1985 revealed that in less than 3% of reported cases was the photofit "of assistance to the investigating officer" (Bennett, 1985) and underscores the notion that photofit requires the use of supporting evidence to be of any general value.

So, why use a visual representation in the first place? The main reason is that recognition performance *can* be near ceiling level (even when the image quality is poor) if the target is known to the person carrying out the identification (e.g. Bruce, 1988; Burton, Wilson, Cowan & Bruce, 1999; Hancock, Bruce & Burton, 2000; and Koehn & Fisher, 1997) and is

⁴ Although the Identikit II system arguably produces more realistic composites, since photographic elements are used instead of line drawings, no formal analysis of the Identikit II is known (Shepherd & Ellis, 1996). Note that research does indicate the benefit of more realistic representations to face recognition (e.g. Davies, 1982, 1983b; Leder, 1996; and Perrett, Benson, Hietanen, Oram & Dittrich, 1995).

therefore considerably higher than that obtained via a verbal description (Christie & Ellis, 1981). Indeed, that Christie & Ellis (1981) found less than 50% accuracy in matching via a description does suggest people have difficulty in describing faces anyway. Note that this same level of accuracy has been reported elsewhere (Shepherd, Davies & Ellis, 1978). A second reason is that the visual memory of a face does not decay as fast and is known to be more robust against interference than a verbal one (e.g. Davies, 1983a; Davies, Ellis & Shepherd, 1978; and Ellis, Shepherd & Davies, 1980).

Christie & Ellis (1981), Ellis et al. (1978a) and Laughery & Fowler (1980) attribute their poor results to limitations in the composite system, including the limited number of features available. This can be illustrated with, for example, hairstyle. Despite the vast variation in possible hair colouring and style, in general there are only 204 different hairstyles in the original Photofit system and 130 in the Identikit system. The limited number of hairstyles is of particular interest due to the established role of hair in unfamiliar face recognition (e.g. Ellis, 1986).

Other concerns have been expressed by the lack of a decline in photofit performance when composites are created after much longer periods of time after target exposure. Davies, Ellis, & Shepherd (1978) found that their subjects' recognition ability to a target face deteriorated significantly after a period of 3 weeks. However, there was no significant difference in the rating scores between composites made immediately (after a 15 second exposure) or after 3 weeks. The lack of deterioration in performance was suggested as a deficit in the Photofit system.

It is interesting to note that McNeil, Wray, Hibler, Foster, Rhyne & Thibault (1987) also found no significant decline in performance with Identikit composites constructed after 3 weeks. In another Identikit study, Green & Geiselman's (1989) subjects constructed composites after a 15 second exposure to a target. Other subjects attempted to select the target face from a 6 item photo spread using the composite. They found composites were identified at chance level after only a week's delay. In fact, this later piece of research also found chance level of performance for composites rated as being salient or distinctive. Such a finding is contrary to research suggesting that distinctive faces are better remembered (e.g. Hancock, Burton & Bruce, 1996; Shapiro & Penrod, 1986; and Valentine & Endo, 1992). As an example, Light, Kayra-Stuart & Hollander (1979) report a significant increase in accuracy (an increase in hit rate and a decrease in false alarms) on a recognition task for distinctive faces compared with more average looking exemplars. As the memory of distinctive faces is better, one would expect higher quality composites that would be better identified than composites constructed from more typical faces. The lack of such an effect in Green & Geiselman's study indicates yet another deficiency in the composite kit: the inability to represent faces which are atypical or salient.

Several attempts have been made to improve the relatively low matching and rating scores of composites (e.g. Christie & Ellis, 1981; Ellis, Davies & Shepherd, 1978b; and Ellis, Shepherd & Davies, 1975). In one successful study, Ellis, Davies & Shepherd (1978b) investigated the effect of feature demarcation lines present in Photofit constructions. They found that these lines reduced the identification rate (hit rate minus false alarm rate) in recognition tasks where subjects were required to report whether a face had been seen previously. This suggests that Photofit feature boundaries interfere with face processing.

Another study explored the limited number of features available in a composite kit. Gibling & Bennett (1994) had experienced operators construct composites using Photofit with the original photographs as a reference. These composites were then subjected to a standard method of artistic enhancement using acetates that both removed the presence of the photofit feature lines and added elaborative detail. Subjects were given the composites to pick out the targets from a 12 person photo spread. A 15% hit rate was found for the untouched Photofits but this increased significantly to 54% with acetate enhancement (the number of false alarms was found to reduce with enhancement as well). It would appear then that photofits can be more recognizable if enhancement techniques are employed.

The Electronic Systems

An alternative to artistic enhancement might be to increase the number of features available in the manual systems. This has been a consequence of the electronic variants. There are many such systems available to Police Forces globally. Examples include CAD/C, Futon, Mac-a-mug Pro, WHATISFACE, EFIT and PROfit (originally CD-FIT). Sadly, little research has been performed to date to establish their effectiveness. It would appear that most research has been carried out on the Mac-a-mug system (though data is now emerging for the EFIT system).

The Mac-a-mug Pro (hereafter *MAMP*) runs on the Macintosh computer and contains palettes of facial features that an operator can select and assemble. The number of possible composites that can be constructed is about two orders of magnitude greater than the manual Photofit system (data extracted from Cutler, Stocklein & Penrod, 1988). Unlike the manual systems, features can be resized, moved and oriented in a freehand way. An additional Macintosh paint package, such as MacPaint or MacDraw, is available for elaboration work. Theoretically then an infinite number of faces can be created with the MAMP. The absence of the feature boundaries, together with an increase in the range and manipulability of features, plus the presence of a paint package, should in theory result in a high level of performance for this system.

The first paper published to evaluate the MAMP was Cutler, Stocklein & Penrod (1988). In this study, an experienced operator created composites of 10 target faces. Each composite was made with the target in view and operators attempted to create the best

resemblance possible. In one of the recognition phases, subjects were shown the resulting composites and had to select which were the original photographs from among 60 distractors. The identification rate (hits minus false alarms) was 49% and this high level of performance led the authors to suggest that the MAMP system can perform well.

The identification rate found by Cutler et al. is comparable therefore with the figure of 54% found by Gibling & Bennett (1994) and suggests that performance of an electronic system is able to match that of a manual system when used with artistic enhancement. Caution must be applied in making generalizations from this study. Composites were made with the target in view, constituting a non-ecologically valid operational procedure.

In a later study though, Wogalter & Marwitz (1991) had 54 subjects create multiple photofits using MAMP given a short (8 second) exposure to a target. All composites were later reworked with the target brought back into view to examine any decline in performance when facial memory was employed. A further 5 subjects decided which target each of the composites were based and five more subjects rated each composite for similarity. They found a significant increase in both matching and rating scores when composites were created “in-view” compared with memory (and these measures significantly increased as each subject created more photofits). It is interesting to note that there was a good overall matching ability of 40% for composites created from memory. This figure compares favourably with the matching scores of composites created with the target *present* in Gibling & Bennett and Cutler et al. It is believed that the reason the matching scores were so similar is due to the nature of the task. The lack of any distractors in the Wogalter & Marwitz study means that there were only 6 faces in the target array. Subjects could have matched with relative ease on a few facial features, inflating the matching scores.

In contrast, two MAMP studies have examined composite performance in the presence of distractors. Koehn & Fisher (1997) had subjects create composites with the MAMP 2 days after a short target exposure (of several minutes⁵). Despite care in employing techniques believed to maximize performance – including minimal verbalization of face, use of a guided memory technique (in which subjects are encouraged to recreate the context of the event), trait encoding⁶ and assistance of an experienced operator – they found that subjects could only identify 4% of the original faces from a target array containing 5 distractors. They attribute this performance to construction under memory conditions, since composites created in-view by the experienced operator resulted in 77% recognition. This does indicate very poor performance for the Mac system.

⁵ The exact time was not specified in the paper.

⁶ There is evidence though that trait encoding may not be the best method to adopt. For a feature-based composite system, it is likely the case that a more componential encoding would be better (Wells & Hryciw, 1984).

The result from Kovera, Penrod, Pappas & Thill (1997) is just as damning. In this study, 10 subjects created photofits of 5 faculty members and 5 previous high school classmates. Other college students who knew these individuals were unable to identify the composites above chance from an array of 40 distractors. Also, further subjects who were not familiar with the targets were unable to pick them out in a 5 person photo-spread (the composite was displayed to the subjects for 30 seconds before presentation of the photo spread). It should be noted that composite creation was carried out with the use of an experienced operator and there was an opportunity for photographic elaboration in the MacPaint package (the importance of which has been highlighted by Gibling & Bennett, 1994).

It can be seen then that the MAMP system appears to be problematic in the more forensically important condition of composite creation from memory. This observation has been seen with several other photofit systems, including CADC and EFIT. The CADC⁷ system contains a digitized version of the elements present in the Photofit system. It offers the ability to move and resize features in a freehand way (rather like the MAMP). Once again, CADC can overcome concerns regarding the limited number of faces possible in Photofit. Functions are also available to combine and manipulate multiple hairstyles and a blending routine ensures that the feature delimiting lines, characteristic of the Photofit system, are removed.

The CADC performance was compared with the Photofit system itself by Christie, Davies, Shepherd & Ellis (1981). Subjects used either the Photofit or the CADC system to create a composite after a 1 minute target exposure and then with another target in-view. No significant differences in the overall identification rate between systems were found for a matching task with 18 distractors, except for subjects whose first attempt were composite constructions from memory, then there was a significant improvement for the CADC system (from 18% matching accuracy with Photofit to 28% for CADC). Comparing initial constructions from memory, it can be seen that the Photofit matching accuracy of 18% is comparable to that found by Ellis et al. (1975), but CADC's performance at 28% is more impressive. It demonstrates a marked increase in success for an electronic system in a more ecologically valid situation. Sadly, no additional studies using this system are known and therefore it is unclear whether the CADC results are reliable.

More reliable data is emerging from the EFIT system though. EFIT is currently in use by Police Forces globally, it runs on a PC and is similar to MAMP and CADC in that facial features can be selected, resized and manipulated in a freehand way until an acceptable likeness has been reached. Selection of these features is achieved through the use of verbal descriptions arranged in a standardized coding system: the 'Aberdeen' Index (Davies,

⁷ CADC is acronym for the Cambridge-based lab (UK) that created the photofit software: the Computer-Aided Design Centre.

Shepherd, Shepherd, Flin & Ellis, 1986). Standard paint packages, e.g. Adobe Photoshop, can be used for artistic enhancement.

Curiously, despite its widespread use, there are only a handful of papers published evaluating EFIT. Arguably the most useful is a comparative study with the Photofit system by Davies, van der Willik & Morrison (2000). Twenty-four subjects created a composite using both systems, first from memory and then with the target present⁸; 2 targets were used: one familiar and one unfamiliar. A further 24 subjects, who were familiar with the targets, attempted to recognize them and then match them with the original photographs. Overall, composites were recognized and matched better if the target was present and if the target was familiar to the subject during construction. Matching accuracy was 63% from memory, though this was carried out without distractors. Compared against Wogalter & Marwitz (1991), who also performed matching without distractors, the EFIT system is about 20% higher on this measure and suggests that EFIT is superior to Mac-a-mug.

There is only one known direct comparison between the EFIT and the Mac-a-mug system. This formed part of Christine Koehn's Ph.D. and only the abstract is available for inspection (Koehn, 1996). To quote,

"A comparison of E-FIT and Mac-A-Mug Pro composites demonstrated that E-FIT composites were of better quality than Mac-A-Mug Pro composites. However, neither E-FIT nor Mac-A-Mug Pro composites were useful for identifying the target person from a photograph lineup. Further, lineup performance was at floor level such that both E-FIT and Mac-A-Mug Pro composites were no more useful than a verbal description" (page 4640).

This provides further evidence for the limited advantage of EFIT over MAMP, but underscores the poor performance of both systems. The other important message from the Davies et al. study is that there was no significant advantage for the EFITs over the Photofits (for naming and matching tasks), except for familiar targets created with the target in-view. This contrast is of little forensic relevance since composites are created from the memory of an unfamiliar person in a real case. This suggests that EFIT is no better than the older Photofit kit when tested appropriately. Overall, the average naming rate of composites created from memory was also low (17%).

Sadly, even lower EFIT identification rates were found by Davies & Oldman (1999). Specifically, the study was exploring whether holding a positive or negative attitude towards a target created with EFIT would have an effect on future recognition. EFITs were initially constructed from memory of 4 famous faces by subjects who either strongly liked or disliked

⁸ The target was re-introduced for the target-visible condition and subjects worked with the operator to modify the composite.

them (the target was re-introduced later for the target-present condition). A significant effect of attitude (higher for the targets that were disliked) was found only for constructions made in the presence of the target. The resulting average spontaneous naming rate was found to be low at 6%. When construction was made with the target present, the average naming rate only increased to 10%. It is conceivable that this lower level of performance [compared with Davies et al. (2000)] is the combined result of using only a few targets and not knowing how many of the original targets were recognizable by subjects⁹. A follow-up study is obviously required to explore this issue.

The only other known published EFIT study is Brace, Pike & Kemp (2000). This study also employed well-known or famous faces as targets. Interestingly, they found a relatively large effect of the verbal description on the subsequent identification of EFITs. Specifically, they found about a 10% increase in naming rate when composites were created by an operator alone compared with composites constructed by the normal interaction process involving another person (a “describer”). Although EFITs were constructed both from memory and with the target present, it is not possible from this study to extract the recognition rate for composites constructed just from the memory¹⁰. The paper does, however, quote a mean of 24.95% for composites constructed via the “describer”. This provides a rough indication of performance that is not vastly different from that found by Davies et al. (2000).

Of the remaining electronic systems, little evaluation has been carried out. Gillenson & Chandrasekaren (1975) examined WHATISFACE. This is claimed (by the authors) to be the first computerized photofit system and can be used by non-artists to produce sketch-like composites. They demonstrate an 81% matching accuracy with a large number of composites (60) created with the target visible; the composites were used to select the original photographs (with no distractors). This measure compares rather favourably with that found by Davies et al. (2000) for EFITs constructed with the target present (83%). As there is no other data available, it is assumed that this system is likely to be no better than EFIT.

Summary

In summary, it appears that the manual Photofit systems can perform reasonably well (as measured by matching accuracy) if artistic elaboration is permitted. Sadly, Gibling & Bennett's (1994) study carried out enhancements with the target present and in itself provides little indication of performance if elaboration is carried out from memory. However, if

⁹ Davies (personal communication) points out that the study did not check whether subjects in the recognition phase actually knew the original celebrities.

¹⁰ This is because subjects were shown pairs of composites to name, one constructed from memory and the another constructed with the target visible. It is not known which composite resulted in the identification (Pike, personal communication).

enhancement of this type serves to allow greater feature expression, then it may be valid to equate the “enhanced” performance of Photofit to that of the electronic systems. However when one looks at the performance first of the Mac-a-mug system, apparently good performance is only observed when a target is present during construction. Certainly the work of Kovera et al. (1997) and Koehn & Fisher (1997) demonstrate very poor results when construction and/or identification occurs from memory with the Mac-a-mug system.

If one compares matching accuracy, Davies et al.’s (2000) data suggests that EFIT is preferable to Mac-a-mug [being about 20% higher than that found by Wogalter & Marwitz (1991)] and is reinforced by Koehn (1996). Worryingly, Davies et al. found a low overall identification rate for the EFITs (17%) that was not significantly different compared with the Photofits when construction was carried out from memory. In addition, this figure is not markedly different from the identification rate of 25% found by Brace et al. (2000) with multiple composites used for recognition. The other EFIT study by Davies & Oldman (1999) revealed even lower naming rates from memory (6%). Even if EFIT is preferable to Mug-A-Mug, and is no different to WHATISFACE, the likely performance via identification rates is at best low in the most valid mode of construction. The data available to date suggests therefore that the Herculean effort gone into computerization is largely wasted, since the older, manual photofit systems appear to perform just as well in the normal operating mode expected by witnesses.

Holistic Notions

A recurring reason given for the failings of the manual and electronic systems discussed so far concerns the method of construction itself (e.g. Ellis, Shepherd & Davies, 1975; Kovera et al., 1997; and Koehn & Fisher, 1997). Recall that in order to build a composite, features are selected from palettes, assembled into a face and then, in the case of the computerized systems, “jiggled” into an acceptable facial configuration. Ellis, Shepherd & Davies (1975) believe that decomposing and scrutinizing a face into its constituent parts is likely to result in interference in the internal representation of the face stored in a witness’s (subject’s) memory. This notion fits well into a large body of data that strongly suggests that faces are perceived as a conjunction of facial features viewed at the same time. In other words, it is the parallel processing of facial features (specific features in a given configuration) that leads to recognition; one might say that faces are perceived *holistically*.

An early study by Davies & Christie (1982) demonstrated that consideration of facial features in isolation to the rest of the face could be problematic. In their study, subjects rated whether a pair of eyes or a mouth was present in a target. The ratings were collected both in isolation and within the context of a face (made from the Photofit kit), and the target was either

present during the rating or it was made following a 1 minute exposure. Analysis of the rating scores revealed high and significant inter-correlations (correlation coefficients were in excess of 0.74) between the memory and target visible presentations in all conditions except ratings performed from memory with features in *isolation*. This suggests that the internal representation of the face is biased towards a complete or *holistic* face rather than by a set of its component parts.

Similarly, Tanaka & Farah (1993) reveal a facilitation in recognition when facial features are recognized in their normal context. Their subjects learned 12 faces and then attempted to recognize the corresponding facial parts when displayed in a scrambled, inverted, isolated and a normal configuration. Recognition was significantly better only for whole faces displayed normally (and was over 10% higher than for the isolated feature presentation). The authors failed to find a comparable whole object advantage repeating the paradigm with computer generated pictures of houses and their “features” (i.e. doors and windows). This provided supporting evidence that holistic effects appear restricted to human faces.

Tanaka & Sengco (1997) further explored the notion of “appropriate” facial contexts. Subjects learned 6 target faces (together with an associated name) and were tested on their ability to recognize individual features contained in those faces. Stimuli for the study and test phase were unfamiliar to subjects and were constructed from the MAMP software. Recognition was performed in faces with an appropriate facial configuration (i.e. same as the target), a facial configuration where the eyes were displaced horizontally (a *new* configuration) or with features in isolation. Once again, they found a significant advantage for the detection of eyes, noses and mouths in the original configuration over an isolated format (a 12% increase in hit rate). Although the new configuration was significantly worse than the original, it was significantly better than the isolated condition. This is an important finding, and suggests that even an “incorrect” facial context (i.e. the new configuration) can be beneficial to recognition of individual features seen previously¹¹. This finding has been reported elsewhere (Bruce, Healey, Burton, Doyle, Coombes & Linney, 1991).

There are several studies, using a similar paradigm of orientation and feature changes, providing further support for holistic facial representation (e.g. Yin, 1969; and Young, Hellawell & Hay, 1987). Another approach though has been to manipulate the instructions given to subjects to encourage different facial encoding strategies (e.g. Wells & Hryciw, 1984; Shapiro & Penrod, 1986 and Sporer, 1991). These studies directed attention towards the

¹¹ The results of studies such as these have been influential in a change of procedure used to create composites (e.g. Davies, Shepherd, Shepherd, Flin & Ellis, 1986). The issue concerns the selection of facial features. Although it is entirely possible for eyewitnesses to select individual features from palettes, it appears advisable for features to be selected in the context of a whole face. That is, even for feature-based methods, improvement can be made if there is a bias towards a more holistic method of construction.

physical aspects of a face (referred to as analytical, componential or feature-based encoding) or encouraged the assignation of character trait (a *holistic* encoding). Bower & Karlin (1974) were the first to examine this effect with human faces. They had subjects identify gender (analytical) or report on a face's honesty or likeability (holistic). A subsequent recognition task was best for honesty judgments, followed by likeability and worst for gender discriminations.

Shapiro & Penrod (1986) performed a meta-analysis on a large number of mainly lab-based studies (128 in total) in facial recognition and eyewitness identification. 19 main variables were analyzed, including factors such as subject age, gender of target, stimulus exposure and encoding instructions. In the 29 studies relevant to encoding, it was found that the hit rate was significantly greater if the instructions orientated a subject to encode a face with a personality trait rather than to locate a facial feature (a feature-based encoding). Similarly, Coin & Tiberghien (1997) investigated 26 studies comparing judgments about physical features or personality traits. They found that in 25 out of 26 studies, a significant increase in the identification rate (hits) for trait encoding was observed; 8 of these were published after Shapiro & Penrod (1986). Research has even found that performance can deteriorate if a feature-based analysis (as opposed to a trait-based analysis) is carried out at the same time as attempting to recognize a face (Berman & Cutler, 1998).

Despite convincing evidence then for a holistic coding scheme for recognition, such an observation may not be pervasive in all face processing paradigms (e.g. Wells & Hryciw, 1984; Laughery, Duval & Wogalter, 1986; and Wells & Turtle, 1988). In Laughery, Duval & Wogalter (1986) for example, subjects studied a target face and then created a photofit using the Identikit system. A follow-up questionnaire examined, *inter alia*, the natural encoding strategy employed. The resulting photofits were then rated for likeness to their corresponding targets (by a different group of subjects). It was found that the highest rated photofits were produced from those subjects who utilized an analytical or feature-based method of encoding rather than a more trait-based (holistic) approach.

Wells & Hryciw (1984) also manipulated the strategy used for encoding. Their subjects viewed a target under either feature or trait encoding. Half the subjects constructed a composite using Identikit while the other half attempted to recognize the photofits themselves from a 6 item photo spread. They discovered that hit rates were best under trait encoding (a 30% increase) but construction was best under feature encoding (a 10% increase).

Both Wells & Hryciw (1984) and Laughery et al. (1986) lend support to the notion that encoding prior to photofit construction is better if feature-based rather than trait-based. This is not surprising of course since all the systems discussed so far are componential. One possible criticism with both of these studies is that they employed the Identikit system and, along with the Photofit kit, may not be sensitive enough on their own (i.e. without artistic enhancement) in order to capture a likeness of sufficient quality. As mentioned above, much better results can be achieved with the electronic systems if the target is present during construction (e.g. Koehn

& Fisher, 1997; Davies et al., 2000; and Wogalter & Marwitz, 1991). A sensible test then would be to compare the electronic photofit performance under differing encoding conditions. Although this manipulation has not been performed, one would expect to observe better performance under feature encoding.

In a naturalistic setting however, there is good reason to believe that individuals will more often choose a holistic over an analytic strategy when there is no expectation of a memory test; Olsson & Juslin (1999) found that a holistic approach was primarily adopted for 64% of their subjects. This is in contrast to Laughery et al. (1986), where subjects were aware of an ensuing memory task, and 62% were found to have adopted an analytical strategy. Taken together, these results suggest that the current photofitting approaches are biased against those witnesses who are not aware that they need to create a composite at a later date (and tend to adopt a holistic encoding) and those who *are* aware of a test and go onto encode holistically anyway. A further consequence is that a system with an exclusive holistic bias may not be the best system for analytical encoders. Overall then, a hybrid holistic-componential photofitting approach may be optimal for a witness.

Parametrized Models

The prior discussion brings into question the exclusive approach of feature-based methods for the 2D representation of a face for the purpose of facial imaging from memory and posits that a method that can allow a holistic representation may be more appropriate. However, in order to represent and manipulate a face, a parametrizable model is required. That is, there needs to be some way of specifying a face via a set of parameters.

Valentine (1999) explains that the notion of a multidimensional similarity space (MDSS) has been a highly influential approach in the representation of stimuli. Research suggests that, with respect to faces, there exists a “typical” facial representation (e.g. Valentine & Bruce, 1986). In his “face space” model, such a representation is assumed to reside at the origin of the MDSS. Valentine justifiably attributes considerable weight to the existence of a face space based on the significant research in facial distinctiveness. The basic notion, as mentioned briefly earlier in the chapter, is that distinctive faces are better recognized than more typical faces (e.g. Hancock, Burton & Bruce, 1996; Shapiro & Penrod, 1986; and Valentine & Endo, 1992). Bruce, Burton & Dench (1994) found that distinctiveness ratings (of unfamiliar faces) were significantly correlated with physical deviations from an “average” face (provided that the effect of hair was controlled). Likewise, multidimensional scaling (MDS) - a technique that establishes relationships between items given similarity ratings - was employed by Johnston, Milne, Williams & Hosie (1997) with pairwise comparisons of similarity rating scores on 36 faces, half of which were distinctive. The resulting analysis revealed that the 18 faces that

were rated as distinctive were distributed further away [in the space created by the MDS] from the 18 that were considered more typical.

Valentine posits that the “face space” framework contains three broad approaches that specify the physical aspects of a face, the psychological aspects of a face or provide a coding via Principal Components Analysis (a similar result is found with neural networks). For each approach, there is an average or prototypical face in the axial centres and movement away from this point provides a code that increases in intensity according to the appropriate metric specified. For example, a psychological model would employ dimensions that map onto psychological variables (Ashby & Townsend, 1986; and Nosofsky, 1986). It would appear that one or more of these dimensions are related to facial distinctiveness. As discussed above, there is considerable research to suggest that faces significantly different from an average face (a distinctive or salient face) enjoy an advantage in face perception. Also, work by Vokey & Read (1992) and O’Toole, Deffenbacher, Valentin & Abdi (1994) find that typicality¹² has separate orthogonal components for both the *familiarity*¹³ and the *memorability*¹⁴ of a face. Vokey & Read (1992) go so far as to hypothesize that these components are reflected in the face space.

In contrast to a psychological model, a physical face space models the physical aspects of the face. Valentine relates this model to Brennan’s caricature generator (Brennan, 1985). In the generator, 169 coordinate points are assigned to the outline of facial features and are connected by line segments. Each of these coordinate points can then be exaggerated with reference to an internal set of coordinates that represent the average facial location. The result is a line drawing that exaggerates distinctive facial features. Rhodes, Brennan and Carey (1987) find that caricatures of famous people were recognized faster (though not more accurately) and rated higher than the original line drawings. The reverse pattern of effect was likewise found for anti-caricatured faces (where the differences from the “average” are reduced). The main point here is that the physical face space comprises of a single dimension for each of the 169 coordinate points contained in the model. Once again, the axes originate from an average or typical face in the centre of the model.

Another example of a physical face space is employed in the relational manipulations possible in the main computerized photofit systems (e.g. EFIT, PROfit and MAMP). The face space model is being explored each time a feature is moved (e.g. moving the eyes closer together). In this case, there would be two main dimensions for each possible feature manipulation, one vertical and one horizontal (although other dimensions can be conceived for feature size, feature rotation and variation in intensity).

¹² *Typicality* can be thought of as an inverse function inverse of distinctiveness.

¹³ A measure based on the degree of confusion between faces.

¹⁴ A measure depicting the ease by which a face may be remembered.

The third type of face space proposed by Valentine concerns the use of Principal Components Analysis (or neural networks¹⁵). Principal Components Analysis (PCA¹⁶) is a statistical technique that can be used for data representation and compression. An underlying assumption of PCA is that there exists a lower dimensional space or manifold in which data can fit. Computation normally involves an initial data normalization (to remove item scaling and bias) followed by the computation of a covariance matrix. The orthogonalization of the matrix results in a set of eigenvalues and eigenvectors. The eigenvectors are produced such that the first one captures the most variance in the data, the second captures most of the variance once the first has been removed, and so on. Re-construction of the original images is possible by a linear weighted sum of the eigenvectors.

The PCA process has overcome what has been referred to as the “curse of dimensionality”. Murase & Nayar (1995) explain that even for a small image database of 100 views of 100 images, this results in 10,000 images (or dimensions) being located in a highly sparse space. PCA would permit a compressed representation in (say) 10 dimensions, a compression ratio of over 1,600:1 (with an image size of 128x128 pixels). Projection to a lower dimensional space then vastly increases image density and also permits intermediate representations. This last point is important since it is the ability to generate new or novel faces (i.e. representations different from those in the database) that is of value should the technique be used as part of a photofit system.

Sirovich & Kirby (1987) was the first study to demonstrate that faces could be represented well with PCA. They started with monochrome photographs of 115 full-face Caucasian males. Simple normalization was performed that aligned the head in the vertical plane¹⁷, the eyes in the horizontal plane and resized the image to make the width of the head the same in each photograph. They found that the first eigenvector (*eigenpicture* in their terms) represented the arithmetic mean intensity of the faces in the set and that the original images could be reconstructed with a good likeness using the first 50 parameters (to within a 4% normalized error¹⁸). This study cropped the images to reveal just the eyebrows, eyes and nose. In later work, Kirby & Sirovich (1990) extended the analysis to include the front part of the

¹⁵ Neural networks are a class of simulation techniques inspired by the morphology and/or function of neurons (Rumelhart & McClelland, 1986). As they have been found to produce results similar to that obtained by Principal Components Analysis (e.g. O’Toole, Abdi, Deffenbacher & Valentin, 1993; and Linsker, 1986), they will not be considered separately in this discussion.

¹⁶ The technique is also known as the Karhunen-Loeve (LV) expansion and the Hotelling Transform and was first described in 1901 by Pearson (Kirby & Sirovich, 1990).

¹⁷ This was achieved by manually aligning each image so that they overlapped about the line of vertical facial symmetry.

¹⁸ This is the RMS error between the original image and the reconstructed image, divided by the vector length of the reconstructed image.

hair, the forehead and the mouth (by presenting the face as an oval cameo shape). Once again, the majority of the variance (about 95%) was captured in the first 50 eigenpictures.

Craw & Cameron (1991) observed that Sirovich & Kirby (1987) and Kirby & Sirovich (1990) used an ad hoc method of image alignment. This was necessary in order to limit blurring effects that occur when interpolating between faces in the face space. These studies performed PCA on the *pixels* in the images of the database corpus. This necessarily means that unless all the facial features are aligned, the pixels defining the facial features in the eigenpictures will not completely overlap and a noticeable blurring effect will result when the eigenpictures are interpolated. The effect of their crude alignment procedure can be seen in the “smudged” appearance around the eyes of the average face (refer to Fig. 1 in Sirovich & Kirby, 1987).

There are other methods of feature alignment (Brunelli & Poggio, 1993; Craw & Cameron, 1991; and Troje & Vetter, 1996). For example, Craw & Cameron (1991) located coordinate points (Craw & Cameron refer to them as *control points*) around the major facial features (eyes, eyebrows, nose and mouth) and the outline of the head including the ears, chin and jaw. The average position of each control point was computed across the image set and the image was triangulated to produce an image mesh. Each database image was then morphed to the common face shape before performing PCA. The common shape was achieved by distorting or *morphing* the areas of the image defined by triangles (a bilinear interpolation) such that all triangles had the same common shape (they refer to the resulting image as *shape-free*). In their study, they repeatedly demonstrate that faces not part of this image database can be constructed to an “almost identical” accuracy using a linear combination of eigenvectors.

Hancock, Burton & Bruce (1996) argue that the control point information itself can form part of a PCA that models the relational aspects of the face (e.g. the distances between facial features). They refer to this process as a *shape* PCA model with the resulting eigenvectors termed *eigenshapes*. The second model, concerning the *shape-free* image intensities, is referred to as the *texture* PCA model (and the associated eigenvectors are referred to as *eigenfaces*). The term *texture* is used in a restricted sense in the paper, referring to the information in the image that remains after the face has been shape-normalized (i.e. made “shape-free”). Reconstruction of a face (or creation of a novel face), begins with a weighted recombination of the *eigenfaces* (i.e. from the texture PCA model) and the *eigenshapes* (i.e. from the shape PCA model), producing a shape-free face and an associated control point vector. The image is then morphed (from the average shape of the image) to a new shape defined by the control point vector. Thus, a fully parametrizable face model is available using these techniques.

It is worth pointing out that PCA used on a dataset as described above naturally produces a global facial representation (Hancock, Bruce & Burton, 1997). Each dimension of the subspace (i.e. the eigenvectors) provides a representation that affects the entire image (rather than an isolated part of it). This is due to the computation of a covariance matrix that associates components of the face that change at the same time. For example, Hancock et al. illustrates

that the first subspace dimension for shape provides a representation that looks as though a head is “nodding”. This said, it is conceived that a more analytical or feature-based approach is also possible with this model. Recall that coordinate points were used to highlight the features of the face. It is entirely possible to “move” these features in a free-form way by simply morphing the image specified by the control points. This simple approach would permit a componential exploration of the face space (rather like the current electronic composite systems) and result in both a holistic and an analytical implementation.

Which Method(s) to Adopt?

Of the three broad approaches in the face space framework proposed by Valentine (1999), all three *could* offer a holistic solution for a photofit system. Intuitively, the most appealing is the psychological one as such a model could represent a coding scheme analogous to that found in human face perception. However, despite indications of the established importance of distinctiveness, familiarity and memorability, the nature of the dimensionality of this space is currently unknown. In contrast, the caricature generator of Brennan specifies the physical aspects of a face and is an inherently holistic approach. However, the generator requires an external facial image in order to create faces and therefore it is difficult to imagine how one would explore the face space without such an external reference. On the other hand, a method involving PCA seems the most promising since it can offer not just a holistic solution but an analytical one as well. Using Valentine’s terminology, this hybrid solution would implicate both a physical and a PCA approach.

Interestingly, there is mounting evidence in the literature that links PCA and face perception (e.g. Hancock, Burton & Bruce, 1996; O’Toole, Abdi, Deffenbacher & Valentin, 1993; and O’Toole, Deffenbacher, Valentin & Abdi, 1994). For example, O’Toole et al. (1993) found that PCA performed on a mixed gender Caucasian database can be used for both gender classification and face recognition; O’Toole et al. (1994) has shown that such a model can parallel human performance in several measures such as typicality, familiarity and attractiveness. Furthermore, PCA has already been applied in a forensic setting to search for targets in mugshot albums (Baker & Seltzer, 1998); refer to Chapter 3 for details.

Towards a New Approach

The most promising route forward in producing a photofit system seems to be based on Hancock, Burton & Bruce’s dual shape-texture PCA model coupled with an analytic-type free-form feature manipulator. This would provide a holistic coding scheme in which, by the very nature of the linear PCA subspace, can be used to generate a potentially infinite number of interpolated faces. Arguably, the easiest method to explore the face space is to directly alter the coefficients of the Principle Components (PC) for both shape and texture. Such an approach

has been attempted by Brunelli & Mich (1996). In their prototype identification system called "Spot It!", they provide a set of slider controls for a PCA performed on each facial feature. A constructed image is displayed for the combined set of slider settings along with an ordered set of mugshots that most closely match the constructed image. The effectiveness of the system is unclear and the authors report that the system is awaiting field test.

Sadly, the PCs appear to exhibit generally complex representations. In his most recent model of a database of 20 Caucasian female faces, Hancock (2000) shows that the first eigenshape produces a "nodding" motion and the second changes face width. The third eigenshape is more complex, having both a rotational component (head tilt) and one that differentially changes the width at the top and bottom of the head. The other components have even more complex behaviours. It would appear therefore that the direct manipulation of the PCA space might not be ergonomic and it is unclear therefore how the Brunelli & Mich system might perform.

The problem of complexity is confounded by the observation that the PCA can generate a very large (and potentially infinite) number of faces, as mentioned above. If one makes the assumption that a suitable representation of a target exists in the PCA space, then conducting an exhaustive search for it is likely to be costly in time. This is largely due to feedback being required from a user to indicate the "quality" of each representation. Overall, this approach appears too impractical to be used with a witness.

There are a set of techniques however that have been developed over the last 20 years to explore potentially complex manifolds like the PCs' under consideration. These come under the umbrella term of "Genetic Algorithms" (or more simply, GAs). GAs generally model processes that occur in nature such as the "mixing" of genetic materials from "parents" to provide one or more "offspring". The GA approach is essentially a parallel one: a search of the problem space is carried out in multiple "places". In a typical scenario, a large number of initial solutions are proposed and a "goodness" value is derived for each. A selection function operates such that the better individuals have a greater chance to take part in "mating". Pairs of these "successful" solutions are taken and the components from which they comprise are mixed (cross-over) to generate a new solution (an offspring). Breeding continues until the previous population size is reached. An evaluation function is again applied and breeding continues as before. The whole process repeats until a further objective function is satisfied such that either the population as a whole or a single individual is of "sufficient" quality.

This is a common procedure but there are numerous other techniques within GAs for exploring problem spaces. For example, it is possible for a small group of individuals to compete against each other in a "tournament selection" or breed with themselves via asexual selection. There are also many parameters within GAs open to manipulation. Mutation rate, the manipulation of a parameter under the influence of a noise source, is one. On one hand, one could explore the PCA space using a few exemplars. In the limit, this was carried out by

Brunelli & Mich (1996) with a single image and is rather like the impractical exhaustive search method mentioned above. However, one general observation is that the more individuals there are in a population, the greater the chance of finding an acceptable solution. This suggests that Brunelli & Mich (1996) might not have adopted the most efficient approach using a single face.

An approach that uses a relatively larger number of faces in a population has been developed by Hancock (2000); the number being limited to eighteen by the physical constraints of the computer monitor. In his prototype system, a small shape and texture PCA model was built (as described above) from 20 female faces. Eighteen novel faces were generated (with components that were generated from random numbers) and displayed on a computer monitor. Each face was associated with a slider that could be adjusted (by the computer's mouse) to indicate "preference", corresponding to a "fitness" rating between 0 and 10. The program would then select those faces with the higher rating (fitness proportional selection) as parents. The parameters from each offspring face were picked at random from either parent (uniform cross-over) and a small mutation rate was applied to the combined parameters. The author explains that the system is in the early stages of development and evaluation of the general approach is required.

A photofit approach using a GA has already demonstrated good performance by Caldwell & Johnston (1991). Their method was to create a population of 20 faces assembled from selected components from the Photofit kit. As in Hancock (2000), parents were identified by fitness proportional selection, only this time based on a 9 point rating scale (resemblance to a target), and a GA bred another population with uniform cross-over and a small mutation rate. Selection continued until an acceptable likeness was reached. The paper reports that subjects constructed a composite after viewing a simulated crime and "subjective evaluation" was carried out by independent judges. Sadly, their paper is rather limited regarding the experimental procedure and the results obtained, although they do illustrate one composite created after 10 generations (Fig. 4) that appears to have a good likeness with the target. No other known evaluation appears to have been carried out.

A related procedure was adopted by Rakover & Cahlon (1989). Subjects were shown 100 pairs of Photofit faces along with a target. For each pair, subjects selected the one that appeared most similar to the target. A composite was constructed from the features in the faces that were chosen most often. They found that subjects could create composites with about 80% of the features correct. This figure rose to 100% if the data from their 30 subjects was combined. At present, it is unclear how the results would be affected had the composites been created from memory.

But, how can one be confident that faces similar to a target will be responded appropriately in such a parallel presentation? Clearly the utility of similarity rating scales employed in several of the above systems is unclear (Brunelli & Mich, 1996; and Caldwell & Johnston, 1991). However, in addition to the positive result found by Rakover & Cahlon (1989),

there is evidence from other systems that have demonstrated considerable benefit in the selection of whole faces from a presented set (e.g. Baker & Seltzer 1998; and Levi, Jungman, Ginton, Aperman & Noble, 1995). Details of these mugshot-based applications will be discussed in Chapter 3.

Arguably the most compelling evidence of human abilities to select similar looking faces comes from cases in criminal law concerning proven wrongful conviction. Rattner (1988) has carried out a survey of 205 such cases and reports that mistaken identification (as opposed to other causes such as perjury or negligence) took place more than 50% of the time (Sporer, Koehnken & Malpass, 1996). Other, more direct evidence for the appropriacy of facial similarity judgments emerges from a cluster of studies that have reported confusion between faces during perceptual tasks (referred to as the “familiarity” dimension in Vokey & Read’s (1992) work).

It is clear from lab-based research that any confusion between faces is not distributed randomly but met with a high degree of agreement among subjects (e.g. Davies, Shepherd & Ellis, 1979; and Goldstein, Stephenson & Chance, 1977). For example, Davies, Shepherd & Ellis (1979) had subjects sort faces into piles of similar faces. They carried out a multidimensional scaling analysis (HICLUS) and found that in a subsequent identification task, selecting faces in an array from memory, foils drawn from similar clusters resulted in higher misidentifications than foils drawn from different clusters. Indeed, the false alarm rate of common clusters foils accounted for over 70% of the errors, indicating a high degree of agreement across subjects. On the other hand, Laughery, Fessler, Lenorovitz & Yoblick (1974) selected foils based on either similarity ratings or physical similarity (more features in common). In either case, the ability to recognize a target in a sequential search task decreased when more similar foils were employed. Likewise, Courtois & Mueller (1981) found that the false alarm rate was significantly higher if both the target and foils had been previously rated as “typical” (as opposed to being rated as distinctive). This last study fits into a larger body of research (mentioned previously) suggesting that distinctive faces are better recognized (e.g. Shapiro & Penrod, 1986). Therefore, that subjects are able to confuse similar items (obtained by ratings, clustering algorithms or by virtue of their typicality), suggests that they should be able to identify those items that are similar. Such a hypothesis does of course require verification in the application of a face evolver.

Despite the advantage of a more recognition-based approach, one problem with the use of the electronic photofit systems (adopted by Caldwell & Johnston and Rakover & Cahlon) is that it is possible to create composites that do not appear “very realistic” – as has been reported (to me) by photofit operators. The problem here is that it is possible to produce a face with unusual spatial relationships: for example, a face with an eye subjectively too high in the face or a mouth that is implausibly wide. As Ellis & Shepherd (1992) demonstrate (in Fig. 3.8), one can position facial features in arbitrary positions. Although Ellis & Shepherd’s example is

extreme, a system that does not inherently permit “configural extremities” is considered valuable.

It is believed that the holistic model built by PCA would guard against such inappropriate effects. This is because the spatial relationships between facial features are based on the statistical variation of faces in the database. Sampling points in this face space then results in a novel face with plausible spatial relationships. Indeed, that these relationships are not specifically modelled and/or constrained in the electronic photofit kits is potentially problematic.

Conclusion

In conclusion, research suggests that the manual photofit systems produce poor quality composites and the computerization thereof appears not to have been an improvement in a “forensically friendly format” (i.e. constructions carried out from memory). A major problem is their analytical nature, given that face perception is inherently holistic and some people will tend towards a holistic facial encoding anyway. A holistic approach would therefore not only more closely match face perception but guard against the apparently strange spatial relationships achievable by the electronic composite systems. The ability to manipulate facial features voluntarily would also permit a more feature-based approach observed in some individuals. Thus, a dual shape-texture PCA model with a GA front end and a feature manipulation utility is believed to be the answer. This approach may overcome the “failure of composite systems to capitalize on the witness’s recognition abilities” (Davies, 1983a, page 117). The following chapters develop an implementation.

Chapter 2: Pilot Work (Mark I Face Evolver)

It is unclear at present whether a holistic face model based on Principal Components Analysis with a Genetic Algorithm as a user interface is likely to be successful as an approach to implementing a photofit system. This section considers appropriate hardware architectures and software tools as an investigative framework for developing this kind of photofit system. Ultimately this work leads to a simple design that models only changes in image intensity (referred to as a shape-free or texture model) for a small database of faces. It will be seen that even for this "minimal" system that the design considerations are considerable. Ultimately, the pilot software (also referred to as the Mark I Face Evolver) looks very promising and several design improvements emerge, particularly concerning the use of rating scales and parental sampling, which serve to promote better implementations in further chapters.

Design Considerations

A crucial decision early in a project is the selection of appropriate equipment and tools. It has already been decided that a model using Principal Components Analysis (PCA) and a Genetic Algorithm (GA) be considered as key players in an initial solution to a new photofit system. As the operations necessary to generate images from a PCA model alone are too time consuming to be performed manually, the proposed approach lends itself to a computerized solution. Two additional design decisions naturally arise. The first involves the choice of computer system and the second, the selection of software that will run on the chosen computer.

Computer System

The choice of hardware in any project is likely to be important. If ultimately a solution becomes tractable for a photofit system, then it would be highly advantageous to have the "final solution" in a format that can operate directly on a user's computer. One would want to avoid a large and potentially costly exercise in the re-design of a system. This does not of course exclude the possibility of creating solutions for other computers, although this is perhaps preferable after a system has been accepted.

There are a number of computer systems that could conceivably be used to run photofit software. Several have been mentioned already in the last chapter. These include a P.C. (e.g. currently running EFIT and PROfit), a Macintosh (running Mac-a-mug) and architectures that run the UNIX operating system (e.g. a Sun or an Apollo - used to develop Peter Hancock's prototype Face Evolver). All these systems are currently available to the police. It is generally accepted that a P.C. is the most widely used and generally pervasive computer in

the world. The observation that two of the leading photofit systems in this country (EFIT and PROfit) run on a PC is significant. In addition, the electronic version of the Identikit system also runs on a P.C. (Smith & Wesson, 1997) as do other systems available in America: Compusketch (Visatex, 2000), SuspectID (ImageWare, 2000) and comPhotofit (Sirchie, 2000). If it can be shown that a holistic approach can work, then it is sensible to have an initial implementation running on the same system as the majority of others. For this reason, a solution involving a P.C. appears to be the best.

Software

There are two main areas of choice for software selection. The first involves the type of Operating System and the second the programming languages used to create a software solution. An Operating System (O.S.) is the main software that runs a computer and allows other software packages to be executed. It also provides file management facilities, user accounts and peripheral control (e.g. enabling a keyboard and a mouse to be used). There are many O.Ss to choose that will run on a P.C. At the time of starting this project in 1998, Microsoft offered D.O.S., Windows 3.x (e.g. Windows 3.1 and 3.11), Windows 95 and Windows NT. In addition, there is IBM's OS/2 available that can also run on a P.C.

Arguably, Windows 95 and Windows NT are the most common O.S. for a P.C. Once again, PROfit and EFIT will run under these two systems, making either O.S. a sensible choice. Theoretically, it is the case that the same windows program will run under either system (or at least that has been a Microsoft design consideration). An essential difference between the two is that while Windows 95 has the same design philosophy as Windows 3.x and D.O.S., Windows NT permits greater stability, making it a preferable environment in which to develop (Kruglinski, 1997). NT is therefore selected as the initial O.S. for the project (verification of design for Windows 95 will be necessary if a product become feasible).

The second consideration is the type of language or languages to be used in the development a solution. When the project was conceived, it was anticipated that *two* software components parts would be required. The first was a user interface, the second, was a manipulatable model containing the programming for the shape/texture PCA and the GA. Although a single program could be used to implement both (as in Hancock, 2000), it was envisioned that the tasks to be performed were quite different.

To fit in with the style of operation common under windows, a windows-based application (aka a GUI or *Graphical User Interface*) would appear to be the most appropriate user presentation. It provides a high degree of commonality between applications. In a typical program, the mouse is used to select options from a menu bar, performing an action, often displaying another (*child*) window. The *child* window might contain another menu, with

buttons to perform other actions and *text-boxes* to enter information. In other words, skills that a user has developed in one program can be applied to another and (with good design) can considerably reduce a user's cognitive load. Therefore a software design that promotes a windows-based application would appear to be preferable.

The proposed design involves generating and displaying a set of faces (a "population" of faces) and the subsequent "breeding" within this population (followed by an update to display the "offspring" faces). There is a division of labour here. On the one hand, faces must be displayed and selection information collected, and on the other, a face model must initially be created by PCA and then manipulation of the eigenvector coefficients (as part of the GA) followed by recombination of the eigenfaces is necessary (to create a population of faces). Whereas the former is windows-based, the latter is largely numerical in nature (a "number crunching" exercise). It would appear desirable then to locate computer languages that can facilitate the design of these two tasks.

The computation required to perform a PCA can be written as a set of matrix operations. In this case, a matrix takes the form of a representation of one or more photographs of faces. It is standard to represent photographs in an electronic format as a set of pixel intensities. The texture model will therefore comprise of operations on a collection of pixel intensities, one set per photograph. The matrix operations necessary include normalization - to make each image have a mean of zero and standard deviation of one; computation of the covariance matrix and the corresponding eigenvectors; and extraction of the eigenfaces by multiplying the eigenvectors with the original images. Generation of an image involves adding a weighted sum of eigenfaces to the average face, another matrix computation. A similar procedure is necessary to create the shape model for the feature coordinates (control points) defined for each face.

One computer language that has been designed to operate on matrices is *Matlab*. It can run on a number of different types of computer, including a P.C. Matlab is also a *procedural* language¹⁹ and instructions have been optimized for execution speed. In addition, the instructions can be highly abstract. For example, the instruction *cov* computes the co-variance of a matrix; the manual writing and debug of such a function, even in a high-level scientific language like C, requires a very significant programming effort. For these reasons, Matlab (version 5.1) will be chosen as the language to generate and manipulate population images.

Matlab has a set of software tools that enable GUI design (Matlab, 1997). This includes normal window components (such as menus, buttons, text boxes and sliders) plus visually

¹⁹ A *procedural* language contains explicit instructions for a computer to execute. This contrasts with a *declarative* language, like LISP and Prolog, where the problem is described and the computer "decides" how best to solve it.

based tools to facilitate the development of these components. However, despite the ease in matrix manipulation, it is arguably not the best language to use for GUI applications. This is based on an initial assumption that the programming effort for the GUI is likely to be considerable (involving several orders of magnitude more lines of programming than for the generation of images). For relatively large applications, current programming practices appear to be favouring an Object Oriented (OO) approach (Drozdek, 1996). One of the reasons for this is that OO languages tend to facilitate a more modular type of design (i.e. "encapsulation") that encourages the re-usability of code either through the adaptation of data structures in a different context (i.e. "polymorphism") or by the functionality acquired from their "parents" (i.e. "inheritance").

In OO, one writes a program as a set of *classes*. In this way, each class provides an abstraction to one aspect of the problem. For instance, one could have a class to display a population of faces on a monitor, another to record user selections and perhaps a third to "communicate" with the image production software (e.g. Matlab) to exchange selection scores and collect population faces when ready. Overall, development can proceed more rapidly (than non-OO languages – like Matlab, Pascal and C) and code tends to be more readable mainly due to inherent modularization.

There are many OO languages available: Visual BASIC, Small Talk, Java and C++ are examples. All of these could offer a software solution. There exists a *de facto* industrial standard for writing Windows-based programs in OO: Microsoft Visual C++. It uses the C++ computer language and has a programming environment (the *Developer Studio*) that allows the rapid development of windows applications (Blaszczak, 1997). For instance, one can use the Developer Studio to rapidly design, compile and run a dialog window in C++ to collect a piece of user information (e.g. a rating score). It was envisioned that many such dialog windows would be necessary. One of the benefits of Microsoft Visual C++ is that it contains a very large set of classes that are provided as part of the language; the *Microsoft Foundation Classes* (MFC). This is advantageous since they can cut development time considerably by not having to write code from scratch. For example, MFC version 4.2 contains classes for string manipulation, system timing, file management and associated code for windows functions. In fact, the size of this class is over 2MB; a considerable size for a runtime library.

In summary, although many hardware and software tools are available for the project, it would appear best that a solution should be carried out on a P.C. The P.C. should run the Windows NT operating system with Microsoft Visual C++ as a language used to display population faces and collect user input, and Matlab, for the creation of the face models and the generation of population faces.

Pilot Design

It was thought sensible at this stage to run a pilot study consisting of a relatively simple program to evolve a population of faces. This would be achieved by using only the texture model. Given that the programming effort was considered significant to model a full shape-variant face, the effects of facial shape were left for future work. This necessarily means that the faces generated have the same facial shape. This has the benefit of investigating the key aspects of the proposed approach without the initial commitment of a large programming effort.

With this simplified model, a target face could then be evolved. However, this does raise an issue regarding the origin of a target. Basically, the target can be either generated from within the Face Evolver or be external to it. In the former case, a face would be randomly generated from the face model (rather like a face would be produced in the initial population) and act as a target. In the latter, a face that was neither part of the original dataset nor was randomly generated from the resulting model would be selected as a reference image. This “external” target could further be either familiar or unfamiliar to the person creating the photofit. A familiar face could be a friend, colleague or a famous person such as a film star or sports celebrity. The origin of the target will serve to ask different questions. When the target is generated internally, evaluation would examine the ability of the system to “locate” that representation within model’s face space. Externally derived targets test the ability of a system to extrapolate beyond the original dataset or corpus. Whereas internally derived targets can be seen to *always* exist within the model, this is not necessarily the case for the externally derived version and this is therefore a potentially harder task. If the evolution process was found to be not successful with an external target, it would not be clear if this was due to an inability in the search mechanism or a lack of representation (or extensibility) in the face model. For this reason, the initial pilot will explore the use of an internally derived target.

An associated issue is the type of encoding a subject might adopt [either implicitly or explicitly] to remember a target face. As discussed earlier, facial encoding appears to be based on either personality traits or the physical aspects of the face (e.g. Laughery, Duval & Wogalter, 1986). It is hypothesized that the holistic-based photofit system under consideration would benefit from a trait-based encoding method. This is in opposition to current photofit systems, which appear to enjoy the latter, more componential approach when making photofit constructions from memory (Wells & Hryciw, 1984). Importantly, the encoding strategy may be difficult to “enforce” anyway and may be best avoided altogether at this stage. This can be achieved by having the target displayed all the time during the evolution process; an approach that reduces memory load and focuses evaluation on the system rather the user.

The face model was further limited to a small corpus of faces. Now, the shape normalizing process requires the manual positioning of coordinate points around the major facial features and is rather time consuming. Results from Craw & Cameron (1991) suggest that 50 faces are sufficient to build a texture-type PCA model and to create a face that is not part of the original 50 (i.e. an “unknown” or “novel” face) with an “almost identical” approximation²⁰. As the current task does not require this kind of extrapolation, 35 faces were ultimately chosen for the corpus (i.e. a sensible tradeoff between time taken to align coordinates and building a database of sufficient size to capture a good face model). If the approach proves successful, a larger and more realistic database could be assembled.

In addition, the size of the face model was further simplified by discarding colour information. The use of colour information can triple the size of the PCA model²¹ and it is not necessary for face perception (e.g. Davies & Thasen, 2000; Kemp, Pike, White & Musselman, 1996; and Perrett, Benson, Hietanen, Oram & Dittrich, 1995 – but refer to Chapter 6 for a discussion on this issue). Eight-bit grey scale values will be used for each pixel; i.e. an intensity value in the range of 0 to 255.

There is also an issue regarding the demographic profile of faces in the corpus. In this case, what should the gender of the database be? Single gender, or like Blanz & Vetter (1999) and Troje & Vetter (1996), a mixture of males and females? How old should the face be and from what ethnic origin? A report by Gottfredson & Polakowski (1995) reveals that most crime is committed by males in their late teens and early 20s. A design that best fits the profile of the offender is considered most valuable and will be implemented here. Therefore young, male faces will be used. As most participants in this study are likely to be Caucasian and, to guard against potential cross-race effects²² (e.g. Bothwell, Brigham, & Malpass, 1989), the database will comprise of Caucasian faces. Interestingly, the first database in Jacques Penry’s original Photofit kit was Caucasian males (Penry, 1974). Later work could implement further databases.

An associated issue is pose (orientation of the head) of faces in the database. In Shapiro & Penrod's meta analysis, there were 10 studies that investigated recognition ability in $\frac{3}{4}$ view compared with front or profile. Significantly more hits were found for a $\frac{3}{4}$ view compared with a frontal view (33% more) and significantly more hits were found for a frontal view compared with a profile view (100% more). It would appear therefore that a $\frac{3}{4}$ view is the best pose to construct a composite. On the other hand, photofit systems have focused on the creation of full-frontal composites (as opposed to a profile or any other projection). This is an interesting

²⁰ Craw & Cameron (1991) only provide a qualitative analysis of the reconstruction of unknown faces.

²¹ One method is to create a PCA model for each of the three primary colours: red, green and blue (e.g. Perrett, Benson, Hietanen, Oram & Dittrich, 1995). This necessarily increases the size of the face model three-fold.

²² The cross-race effect is a difficulty in identifying faces of a race other than our own.

observation since a full pose may not be best, but there is a move towards such an optimal approach with effort being made to implement $\frac{3}{4}$ view databases in the PROfit system²³. Perhaps the best route for this project then, is to initially evaluate a full-face holistic model. This would provide compliance with other systems and enable comparison to be made against them by keeping pose a constant factor. Later research could investigate the effect of head orientation.

A full-face model also featured in Peter Hancock's prototype system. In that system, the selection of parents was guided by a simple 10 point Likert scale and rating (via a horizontally-oriented windows slider placed underneath the image) was carried out with all faces present. Another method is to rate faces in isolation to the population. This is likely to produce different results. Rating in the context of other faces can provide a frame of reference for comparison. This may encourage a greater use of the rating scale, resulting in a greater selection pressure and faster or better evolution. To investigate the effect of face presentation on rating (and subsequent evolutionary performance), two conditions were tested: sequential face presentation (Condition A) and simultaneous face presentation (Condition B). A simple slider would be positioned under each image for rating.

Constructing a Holistic Face Model

The Department of Psychology at Stirling University has a large database of photographs from the U.K. Home Office. The set contains mainly Caucasian faces of males and females each photographed in profile and full-face pose. The images are available on CD format and can be extracted in a range of image resolutions. Thirty-five of these Caucasian male faces (that were not currently being used for other research in the department) were exported at a relatively low resolution of 300 pixels wide by 400 pixels high (300x400²⁴) and put through a shape normalizing process (morphing) that aligned facial features. As mentioned previously, this stage is necessary to avoid feature misalignment when faces are generated randomly. Forty-two coordinate points were positioned around the eyes, eyebrows, nose and mouth, plus the outline of the head and the four corners of the image. An example can be seen in Figure 1 -

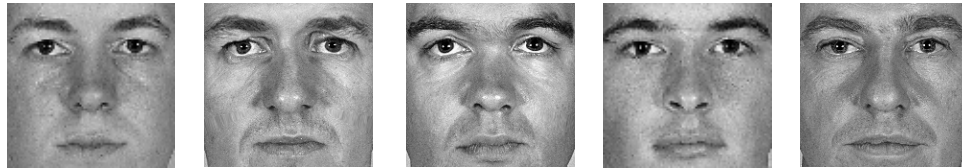
²³ Evaluation of a $\frac{3}{4}$ view database for PROfit is currently being carried out in the Face Perception Lab at Stirling University.

²⁴ The convention of specifying pixel-width x pixel-height for expressing image size will be used throughout the text.

Figure 1: An Example Alignment of Facial Feature Coordinate Points

Following this, each face was morphed to the average shape of the data set and cropped to 176x171 pixels so that just the inner features of the faces remained, serving to limit further misalignment difficulties caused particularly by varying collar-lines and hairstyles. Figure 2 shows the first 5 shape-normalized faces cropped close to the eyebrows and mouth -

Figure 2: Examples of Shape-Normalized Faces used to Construct Face Model



These images were then used to construct a texture model using PCA (in Matlab). Each image was normalized to have a zero mean and unity standard deviation. The covariance matrix was computed for the image set followed by the extraction of the eigenvectors and eigenvalues. The eigenfaces were then computed by multiplying the eigenvectors by the original images. The eigenfaces were sorted by decreasing variance.

Only the average database image plus the first half (17/34) of the principal components (eigenfaces) were used to generate a new or novel face. This was a deliberate design decision since it has been found to introduce error between the database images and the closest approximation generated by the PCA (Sirovich & Kirby, 1987). According to Kirby & Sirovich (1990), approximately 4% (normalized) error should be introduced by limiting the components to this range. This is aimed at providing a level of anonymity for the database images (and is of particular relevance in later versions of the software; e.g. refer to the section titled 'Verifying Anonymity' in Chapter 5). The purpose of the system is to generate novel faces, not the originals.

To produce a novel face, 17 floating-point random numbers (drawn from a uniform Gaussian distribution with zero mean and unity standard deviation) were generated and scaled by the eigenvalue of the relevant eigenface. This has been found necessary to maintain an appropriate influence of each eigenface coefficient (Hancock, 2000). A new image is then the result of the weighted addition of the eigenface coefficients to the average intensity face. A final scaling stage was necessary to produce a consistent image brightness and contrast.

The following experiment serves to examine the performance of a simple evolutionary face generator. The experiment is detailed below but is summarized in Hancock & Frowd (1999).

Experiment 1: Searching the Model

It is proposed that a Genetic Algorithm (GA) similar to that used by Hancock (2000) be implemented. In his model, faces were initially randomly generated and a user would rate how good each was using a slider to define a value between 0 and 10. Similarly, a Likert scale was used in Caldwell & Johnston's (1991) photofit approach. In both studies, the GA would select parents such that faces with higher ratings have a greater chance of being selected as parents for breeding (*fitness proportional* selection). An "offspring" face was composed of coefficients taken randomly from two parents (*uniform cross-over*). The system had the ability to replace or mutate a parameter at random with a small probability.

This general approach was adopted, though it was decided that the range of the rating scale should be from 1 to 10 (rather than from 0 to 10). This would allow all the faces to take part in the selection process even if the rating scale was set to the minimum value (if zero was assigned as a rating, that face would have no fitness value and not be available for selection). This helped to maintain the diversity of faces in a population.

It is also important to decide on a set of appropriate parameters for the system. This includes the number of faces in a population, the number of generations to run and the number of targets to evolve for each subject. A small pilot study revealed that a population of six faces, with 12 generations (the initial set of randomly generated faces plus 11 generations) and five targets would engage a subject for about half an hour. As the rating exercise was reasonably intensive, any longer was believed to be a burden for the participant. It was considered that these parameter settings would likely to result in sufficiently rich data to extract performance trends.

It is acknowledged though that this is a rather small number of individuals for a population. Each member of a population represents an attempt at a solution to a given problem. Relatively larger populations naturally result in more individual solutions, increasing the likelihood of an acceptable solution being found. It is not uncommon for a GA to work with populations containing hundreds of individuals. For example, in Karl Sims's work on evolving creatures that "move" and "behave" in 3D worlds, population sizes of 300 were typical and evolution was performed in blocks of 500-1000 generations (e.g. Simms, 1994). Conversely, with a smaller population size, there is a tendency for faster convergence to occur (as there is less variability in the population). A relatively small population of six individuals then would allow many generations and targets to be used with relatively rapid convergence to a solution.

With such a small population size however, there is the risk that the distribution of solutions within the face space would become too concentrated too early in the evolution process – i.e. the variation of the population would be too low. In an attempt to compensate, a small mutation (that replaces a coefficient with a random number on average approximately once per face) will be used.

Method

Participants

Eighteen students at the University of Stirling participated in the experiment. There were 11 females and 7 males; 9 were assigned to condition A and 9 to condition B. Their age ranged from 18 to 26 and the mean age for males was 21.3 (standard deviation of 2.5) and females was 21.3 (standard deviation of 2.7). They were paid at the rate of £5 per hour.

Apparatus

A Pentium PII PC clocked at 350MHz was used to run the experiment. Faces were displayed on an Iiyama 21" monitor. The PCA model was derived from 35 full-face Caucasian males (extracted at 8-bit monochrome with a resolution 176x171 pixels in BMP format).

Targets

To obtain the targets and to guard against them being too similar, and therefore not exploring the problem or face space thoroughly, a similarity elimination technique was employed. Fifty randomly generated faces were initially produced. A pruning strategy, similar to approaches found in neural networks (e.g. Brown, Hulme, Hyland & Mitchell, 1994; Le Cun, Boser, Denker, Henderson, Howard, Hubbard & Jackel, 1990; and Mozer & Smolensky, 1989), based on the "nearest neighbour" was adopted which repeatedly selected a face at random and then discarded the one from the remaining images that had the lowest mean-squared error. This was continued until only twelve faces remained, resulting in a set of highly dissimilar random target faces. Each face was normalized for equal brightness²⁵ and contrast²⁶ to avoid gross differences in image lighting effects. Seven of these were selected as targets: two for practice sessions and five for the experiment. The first five targets are shown in Figure 3 –

²⁵ Set to a mid-range brightness level of 128 (i.e. a mean image intensity value of 128).

²⁶ Set to a contrast level of 25 (i.e. a value of 25 for the standard deviation of the image).

Figure 3: The Target Faces



Procedure

A further 6 faces were randomly generated (with similar normalization for brightness and contrast) and used as the initial population set to be presented to a subject along with a target face. A different set of initial random faces was used for each trial to guard against idiosyncratic performance. This can occur when one set of initial faces provides either a favourable or an unfavourable set of initial conditions. For example, evolution would tend to be considerably advanced if one or more faces in an initial population were by chance always in relatively close proximity to the target (in face space). Providing a different set of initial conditions for each target would therefore examine more general evolutionary performance.

Subjects were randomly assigned to condition A (sequential presentation) or condition B (simultaneous presentation) and tested independently. They were asked to rate each face for likeness to the target using the slider provided. A demonstration was provided showing that faces more similar to each other were recorded by moving the slider further to the right. An "OK" button was provided to allow rating of the next face(s). No mention of the underlying evolutionary mechanism was provided. All subjects were given a short practice session at the start which involved a single evolutionary cycle for the first two of the seven target faces.

Following rating of the six randomly generated faces, a new population was created. A GA employed a "roulette wheel" mechanism, based on the 6 rating scores, to select a pair of "parent" faces. Uniform crossover (that selected coefficients randomly from either parent) then selected new eigenface coefficients. A small probability of mutation (0.05) replaced 1 in 20 coefficients with an appropriately scaled random value²⁷. The resulting face was scaled for brightness and contrast as before. The procedure was repeated a further 5 times to generate a total of 6 new faces. These became the next population of faces and were presented for further rating. A total of 12 cycles of the evolutionary generator was run for each of the 5 target faces for each subject.

²⁷ Scaled by the standard deviation of the corresponding eigenface.

Rating scores, population faces (eigenface coefficients and actual image files) and demographic information (age and gender) were collected for each subject. Data from the practice session was discarded. Prior to payment, subjects were debriefed with details of the underlying evolutionary system and the Independent Variable (method of rating).

Results

Three measures were used to evaluate performance of the Face Evolver: the mean-square error between the target and the population faces (MSE), rating scores and timing data. Of these, the MSE was used as the primary measure; refer to Appendix A for a discussion regarding this metric. Although self paced, the time taken to complete was analyzed to indicate whether method of rating would naturally affect speed of rating.

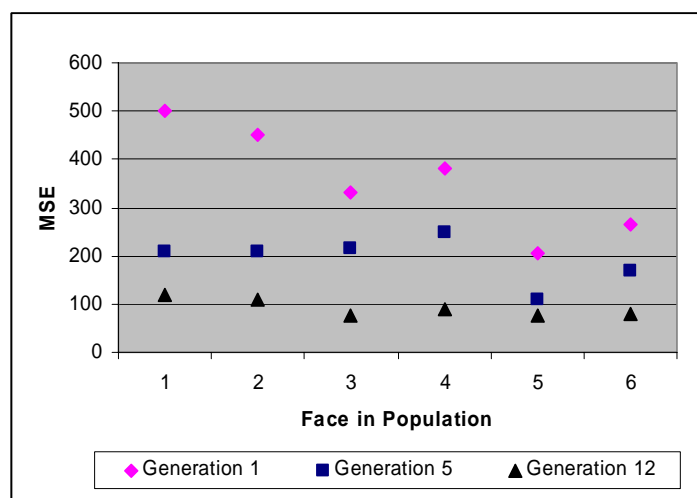
Mean Square Error (MSE)

The MSE measure will first be considered on an individual subject basis, then overall.

Individual (Subject) Analyses

The performance in terms of MSE was obtained for each subject. An example is shown in Figure 4 for Subject 1 on Target 1. The MSE for each of the 6 faces in the population is shown for generations 1, 5 and 12; for clarity, other evolutionary generations are not shown. It can be seen that the MSE scores are generally become lower as the generation increases and is reflected in the average MSE scores.

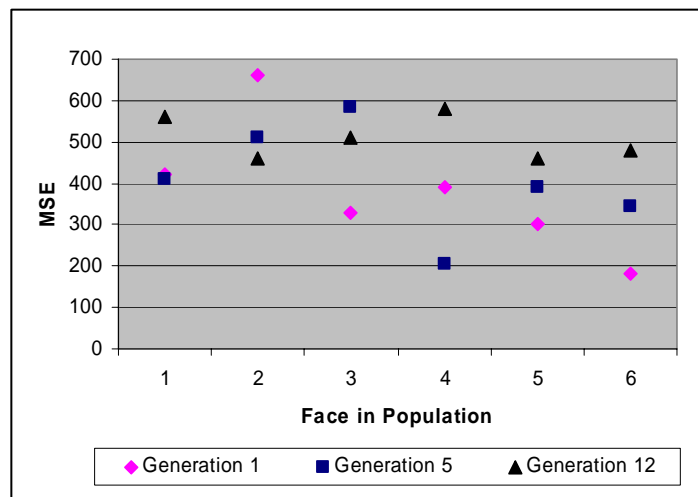
Figure 4: Performance of Subject 3, Target 1



The average MSE in the first generation is 355.0 (SD 112.9). By generation 5 (4 cycles of the evolutionary generator), the MSE dropped substantially to 192.8 (SD 50.6); this reduced further to 89.6 (SD 29.0) by the last generation. A between-subjects two-tailed t-test reveals a significant decrease in MSE at the 0.05 level²⁸ between generations 1 and 5 ($t=3.21$, $DF=10$, $p=0.009$), and also between 5 and 12 ($t=4.33$, $DF=10$, $p=0.001$).

Although these data appears promising, other subjects did not perform so well. For example, Figure 5 displays the MSE for Subject 17 on Target 1. Although the data appear noisier, the MSE does appear to be *increasing* with increasing generation. Indeed, these scores do increase from a mean of 379.8 to 408.1 to 511.1; there is an approaching significant *increase* from generation 1 to generation 5 ($t=2.10$, $DF=10$, $p=0.062$) and a significant increase from generation 5 to generation 12 ($t=3.33$, $DF=10$, $p=0.008$).

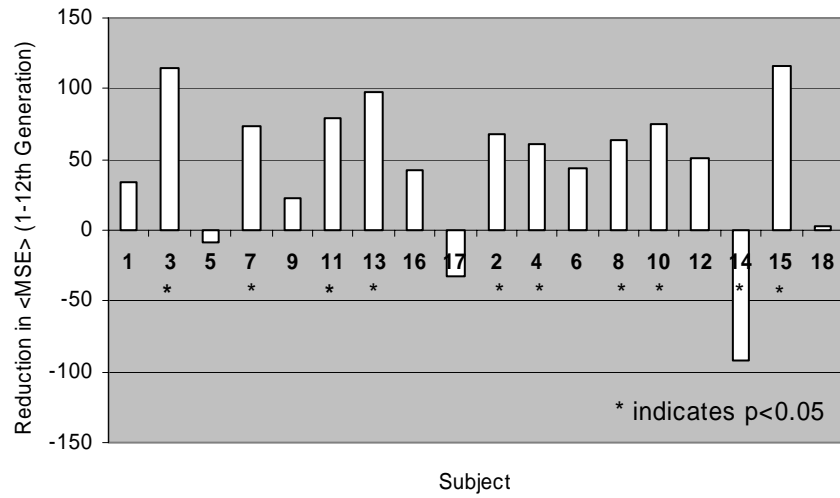
Figure 5: Performance of Subject 17, Target 1



To explore performance further, the average MSE in the first generation (for all targets and subjects) was subtracted from average MSE ($\langle MSE \rangle$) in the last generation (again for all targets and subjects) and is shown below in Figure 6 (for convenience and comparison, the subjects have been sorted into groups A (1, 3, 5, 7, 9, 11, 13, 16 & 17) and B (2, 4, 6, 8, 10, 12, 14, 15 & 18)). It can be seen that 10 subjects exhibited a significant change in MSE between generation 1 and generation 12 over 5 targets ($t>2$, $DF=58$, $p<0.05$).

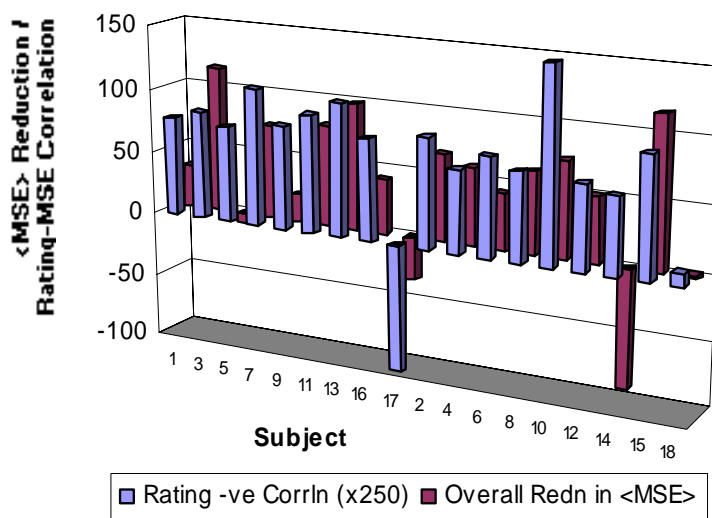
²⁸ A significance level of 0.05 is observed throughout this thesis. Unless otherwise specified, all t-test are “between-subjects two-tailed” in this chapter. The assumption of homogeneity of variance is believed to be upheld. This is not because significant differences in variance may exist (in fact, there is good reason to believe that differences may exist since a population of faces may converge, reducing variance, with increasing generation), but because there is an equal number of subjects in each condition, making the test insensitive to a violation of the assumption (Shavelson, 1981).

Figure 6: Overall Mean Subject Performance (between First and Last Generation)



It is interesting to see a relatively large increase in MSE for Subject 14 and 17. It is possible to gain an understanding of this unexpected increase by examining the correlation between each MSE figure and the associated rating provided a subject; a measure referred to as the *CMR* (the Correlation between the MSE and Rating). One would always expect to see a negative *CMR* since images with higher error scores should be assigned lower values on the rating scale. In Figure 7, the *CMR* has been plotted along with the associated reduction in <MSE>. For clarity, the *CMR* bars (foreground plot) are shown *negative* correlations so as to be viewed in the same sense as the reduction in <MSE> bars (background plot).

Figure 7: Correlation between MSE and Ratings (all Targets)

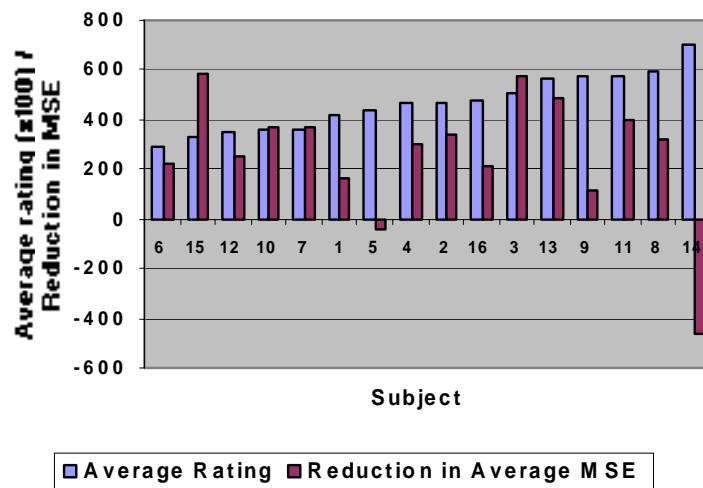


For Subject 17, it appears that the *CMR* is *positive*. It turns out that this subject has rated 4 out of 5 targets with a positive correlation and Subject 18 seems to have used the rating scale

backward for the first target. These participants appear therefore to be using the rating scale backwards. It is proposed that the data from these be removed from further analysis since conformity to the *intended* instructions does not appear to have been followed.

The question arises as to why, despite an appropriately negative CMR, did Subject 14 exhibited a significant *increase* in MSE. Consider a plot of the average rating against the average reduction in the MSE (across the five targets) in Figure 8. From the figure, it is clearly seen that this subject had not only the highest average rating but also the worst error. At this point, it was interesting to see if there was a correlation between these two factors across all subjects.

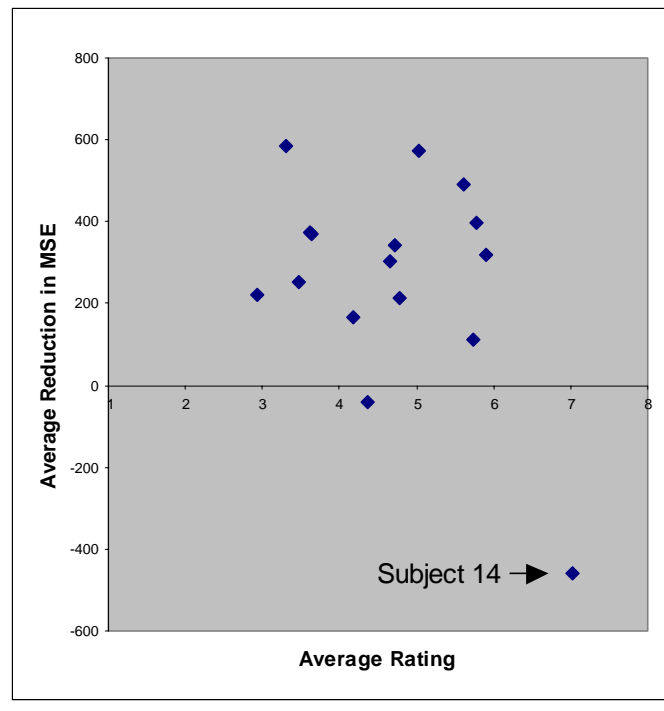
Figure 8: Average Rating Scores with the Corresponding Reduction in Average MSE for Each Subject (the graph is ordered by increasing rating)



Firstly, consider a re-display of Figure 8 as a scattergram (Figure 9). Clearly, there is one data point identifiable as an outlier. This is for Subject 14 and has occurred due to the large *increase* in MSE. Interestingly, ignoring this data point results in a non-significant and near-zero correlation (Pearson²⁹) between variables ($r=0.02$; $F=0.01$, $DF=13$, $p=0.936$) and indicates that there is no relationship between the average rating and the average MSE.

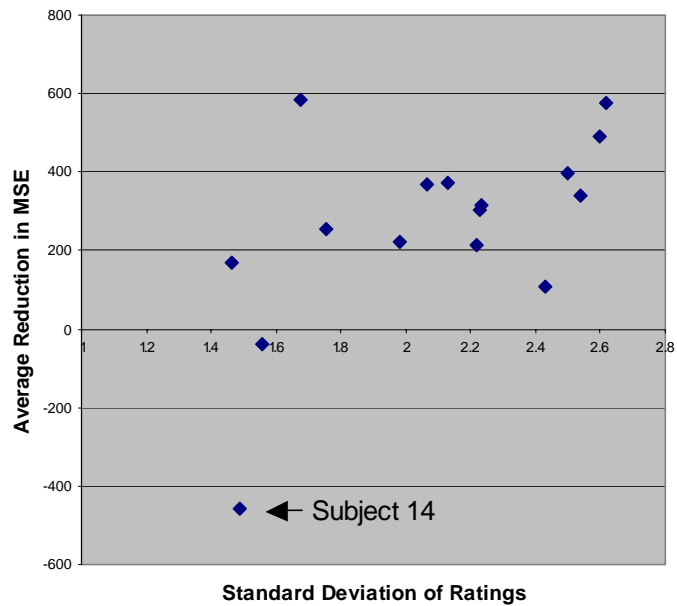
²⁹ Correlation is computed throughout this thesis using the Pearson statistic unless otherwise specified.

Figure 9: No Trend Between Mean Rating and Mean Reduction in MSE



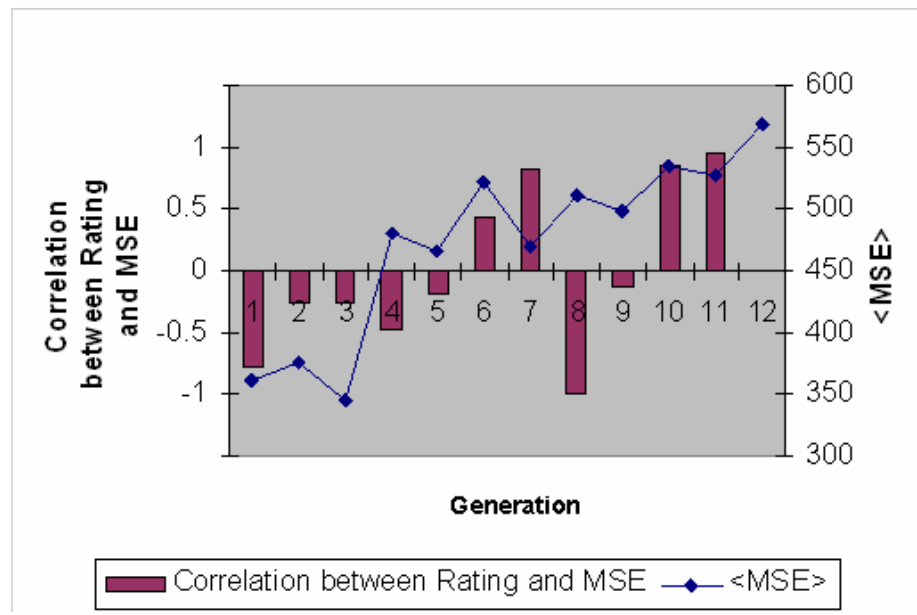
A further interesting finding occurs from the relationship between the standard deviation of the rating scores and the average reduction in the MSE, plotted in Figure 10. Once again, Subject 14 is the only outlying data point. Interestingly, excluding this outlier now results in a medium-level positive correlation that approaches significance ($r=0.44$; $F=3.06$, $DF=13$, $p=0.104$). Note that the positive correlation indicates that greater coverage of the rating scale results in better evolutionary performance.

Figure 10: Positive Trend between the Standard Deviation of Rating Scores and the Average Reduction in MSE



In the light of these findings, the following discussion investigates possible reasons for the poor performance of Subject 14. This will begin by attempting to understand the reasons behind the worst performance of Target 2 for this subject. Consider a plot of average MSE against the CMR for this target (Figure 11). As before, one would expect the trend of a negative-going CMR and a decrease in MSE. However, despite a negative CMR for most of the time, especially concentrated in the first 5 generations, the average error generally becomes worse with increasing generation. In fact, between generations 3 and 4, a significant increase in MSE occurs ($t=2.94$, $DF=10$, $p=0.015$). A likely reason for this is the low correlation between the rating scores and the MSE for generation 3 ($r=-0.25$).

Figure 11: Subject 14's Rating-MSE Correlation and Reduction in MSE for Target 2



Referring to Table 1 below, this resulted in only a small difference in breeding opportunities (0.64) between the best rated face (face 3 with an MSE of 293.15) and the worse rated faces (the other 5 faces all received the lowest rating of 6). But, when sampling for parents, Face 2 (with a relatively high MSE) actually received 4 breeding opportunities out of 6 and the best rated face received only 1. These selections imposed a high MSE on the offspring faces resulting in a significant increase in MSE of 135. An additional effect of this over-sampling of Face 4 was that much of the variability in the population was lost; the SD the MSE roughly halved from 101.1 in generation 3 to 49.9 in generation 4.

Face	Rating	MSE	Breeding Opportunities	Actual Breeding
1	6	463.95	1.89	1
2	6	476.16	1.89	4
3	8	293.15	2.53	1
4	6	228.48	1.89	3
5	6	311.27	1.89	2
6	6	296.29	1.89	1

Table 1: Performance of Subject 14 on Target 2, Generation 3

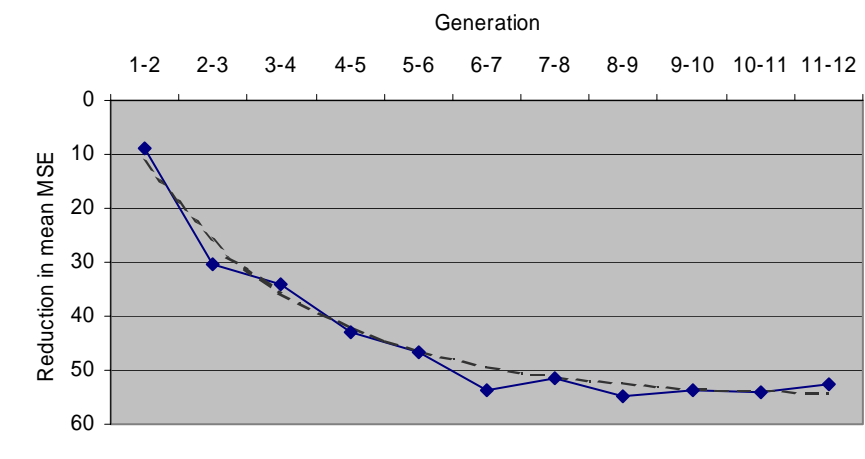
A further example of over-selection of a high MSE population face was also found on generation 3 for Target 5. In general, the results indicate potential problems with using the Roulette Wheel method, where by chance lower rated faces are selected inappropriately often.

This effect is obviously undesirable. The subject has rated as requested (albeit higher than other subjects) but has been let down by the GA due to the selection method employed. Therefore it is reasonable to leave the subject in for rest of the analysis.

General Analysis

Turning now to the average performance for each generation and considering all data (except Subjects 17 and 18). Initially, the reduction in average MSE between each generation across targets was computed. Figure 12 indicates first that the <MSE> decreases with increasing generation. The decrease was largest for the first generation and became progressively less with time, appearing to asymptote on a value of approximately 55. A simple exponential function could be found to describe more than 97% of the variance in the graph³⁰.

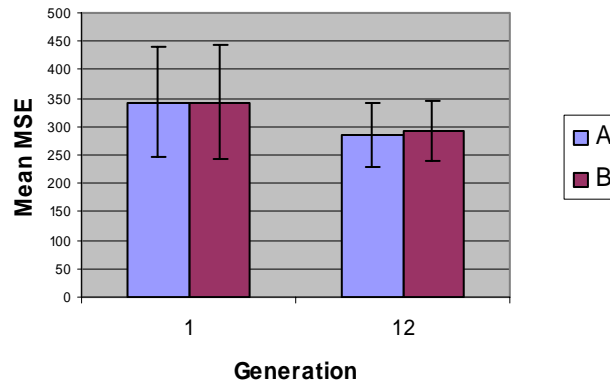
Figure 12: Cumulative Reduction in Average MSE (all Targets)



Comparing performance on condition A (face rating in isolation) and B (rating in the presence of all six faces), it was found that group A had an initial <MSE> of 342.0 (SD 99.9) and a final <MSE> of 294.9 (SD 55.2). Similarly, the initial <MSE> of B was 342.7 (101.4) and the final <MSE> was 299.4 (53.2) -

³⁰ The reduction in average MSE,
 $m = 55 * [1 - \text{EXP}^{-g/2}] - 10$
 and g = generation

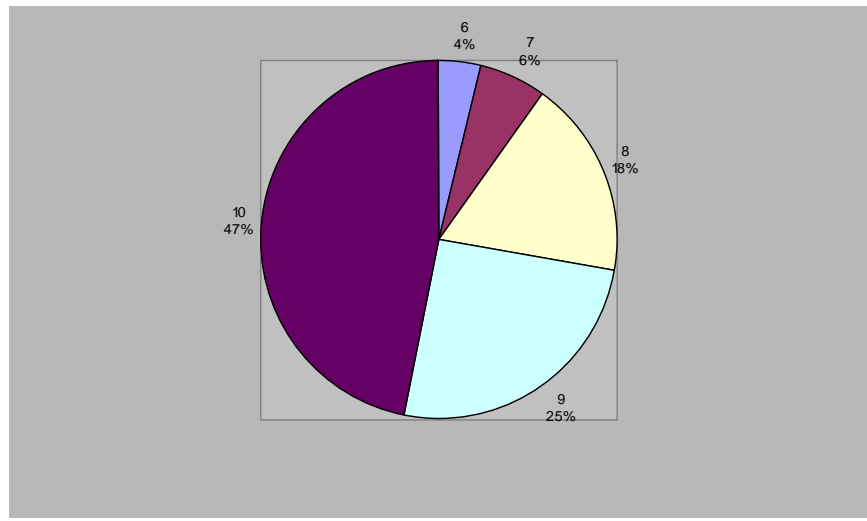
Figure 13: Mean MSE for all Targets (graph bars indicate SD of MSE)



A 2 factor ANOVA was significant for generation ($F=29.79$, $DF=(1,70)$, $p<0.001$) and target ($F=4.53$, $DF=(4,70)$, $p=0.003$) but not for condition ($F=0.08$, $DF=(1,70)$, $p=0.785$). Post-hoc analysis using Tukey HSD for target found a mix of significant differences (between targets) for generation 1 but no significant differences for generation 12. Overall, the decrease in MSE was 45.2 from the first to the last generation.

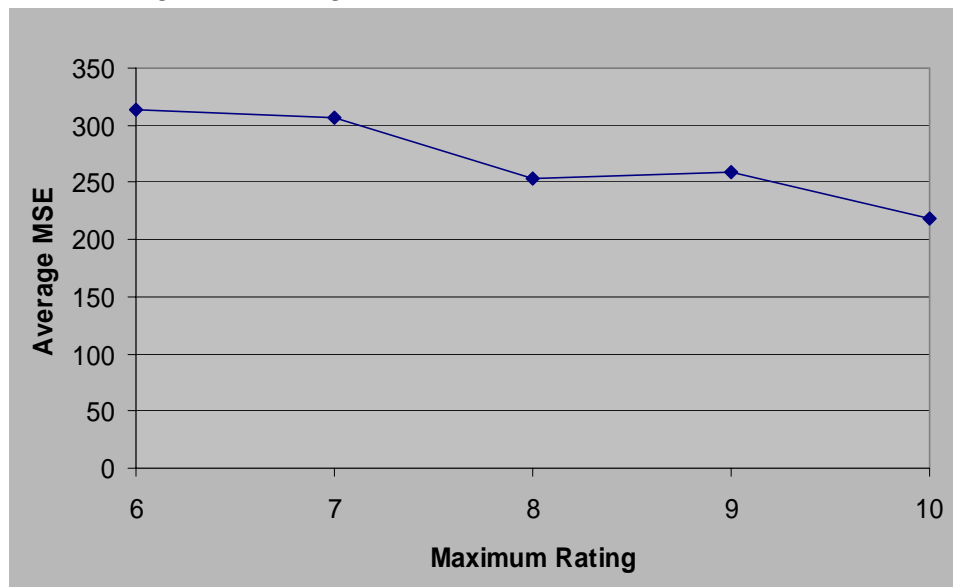
Another approach to evaluate performance is to look at the maximum scores assigned during each evolution, thus providing a measure of peak performance of a target. Recall that there were 5 targets and therefore there will be 5 maximum scores per subject. The average maximum rating is 9.1 (SD 1.1). Figure 14 below illustrates how the maximum rating scores were distributed. One can see that the lowest rating given was a six (there were no observations in the range from 1 to 5). Also, note that the proportion of maximum scores increased with increasing rating category, about 50% of the targets were given a maximum rating and 90% of the targets were assigned a rating in the upper quartile range (i.e. with a rating of 8 or more). The maximum high rating was assigned on average on generation 4.1.

Figure 14: Distribution of Maximum Ratings Assigned For Each Target



The last analysis compares the average MSE scores with the maximum ratings. The average MSE for each maximum rating category has been computed and is shown in Figure 15. Clearly a trend exists, such that as the maximum assigned rating increases, the average MSE decreases; a significant low-level correlation was found ($r=0.28$; $F=6.93$, $DF=79$, $p=0.010$).

Figure 15: Average MSE Score for Faces Rated as Maximum



Rating

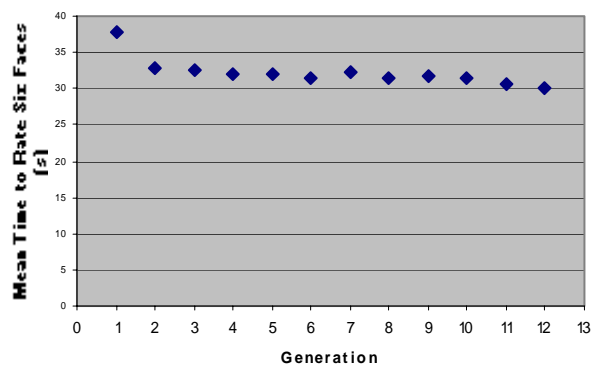
Across all the rated faces, there is a low level significant correlation between the MSE (of the target and a population face) and the subject's rating scores ($r=-0.2$; $F=13.85$, $DF=6478$,

$p < 0.001$). The average rating score for the first generation was 4.15, and 4.76 for the last generation. The mean for group A was 4.57 and for B was 4.35. A repeated-measures ANOVA (using the average rating of each generation to compute a subject's performance) was found to be significant for generation ($F = 14.82$, $DF = (1, 70)$, $p < 0.001$), but not for condition ($F = 0.44$, $DF = (1, 70)$, $p = 0.508$) nor target ($F = 0.76$, $DF = (4, 70)$, $p = 0.555$); no interactions were found ($p > 0.05$).

Timing data

The time taken to rate faces was averaged over subjects and targets. This is shown in Figure 16: subjects were slowest on the first generation and then barely became any faster thereafter. A within-subjects t-test indicates that average ratings in generation 1 took significantly longer than in generation 2 ($t = 8.83$, $DF = 17$, $p < 0.001$); for simplicity, no other t-tests were performed. The mean time to rate six faces for group A was 32.5s (SD 7.58) and 31.9s (SD 10.08) for group B; this was not a significant difference ($F = 0.40$, $DF = 88$, $p = 0.690$).

Figure 16: Average Time Taken to Rate a Population of Faces



Discussion

Data from the pilot work revealed a number of interesting findings. Firstly, two subjects appeared to use the rating scales backwards and therefore not as intended. The experiment had subjects make ratings without any assistance (after a short demonstration). This suggests that supervision is necessary to guard against incorrect use. This is not seen as a problem as photofitting in a forensic setting is carried out with the aid of an operator who is responsible for the correct entry of data.

Operator input can also be seen to be of value even when the scale was used as directed. There is clear evidence that better performance occurred for subjects who used a greater range of the scale (i.e. had a higher rating standard deviation). The reason for this lies in proportional fitness selection used in the Genetic Algorithm. This mechanism assigns a proportionally higher score (referred to as a *fitness* value) to faces with higher ratings. For a set of rating scores (e.g. with rating of six population faces), when the range between the lowest and highest rated faces increases, the ratio of low to high rating increases and the higher rated faces get a proportionally higher fitness values (and the lower rated faces get proportionally less). When the ratio of low to high rating increases, faces with proportionally higher fitness values have a greater chance to become parents. The consequence of which is that “better” faces tend to have more influence the breeding process.

There are two main methods to improve the effectiveness. The first is to “spread out” rating scores mathematically. This can be done in a number of different ways, but necessarily involves transforming the data such that the range increases. One method is apply a linear scaling function³¹. Another is have the operator encourage greater use of the rating scale anyway. The latter may in fact be easier when rating is performed in the presence of all the population faces. In this method of image presentation, the operator could “encourage” subjects to search for the worst and best faces in a population and then guide responses towards the extremes of the scale (with other faces recorded intermediately). In effect, the subject could then “calibrate” their responses.

There is evidence from the data that subjects did not do this naturally though. This is based on the lack of any significant difference in the rating scores between faces rated in isolation (Condition A) and faces rated in the presence of other population faces (Condition B). No extra instructions were given to Condition B subjects that requested them to consider all faces prior to making a judgement. It can be deduced therefore, that parallel rating of faces does not naturally lead to a difference in rating behaviours but *could* be valuable in conjunction with external influences (i.e. an operator).

For the same reason, the use of an operator [or a data transform for the rating scores] could have been of extra value to Subject 14. Part of the problem was that this subject exhibited a low rating standard deviation. The other problem was due to the inappropriate chance over-

³¹ For example, to completely fill the scale from 1 to 10, the following transform may be applied to each rating score (r) for a set of scores (S)

$$R = 1 + [r - m] * 9 / d$$

Where:

m = minimum rating of S

d = (maximum rating of S) - (minimum rating of S)

selection of low-rated faces as a consequence of the Roulette Wheel method of parental selection. Occurring together, the MSE scores tended to *increase* for this subject.

Fortunately, there are other methods available to select parents other than employing a Roulette Wheel mechanism. The problems associated with the Roulette Wheel arise due to the notion of numerical expectancy: that is, it is possible to get over-selection of a random state when few samplings occur, but when the number of samples increase (to infinity in the limit), the resulting random states are more equally distributed. This occurred in this study as only 12 parents were selected for each generation (i.e. 6 pair of faces). Increasing the population size then will, *inter alia*, naturally result in more overall parental samplings and a more ideal sampling distribution.

Another approach of course is to adopt a different sampling strategy that is not sensitive to expectancy. This problem has been addressed directly by Baker (1987). He has designed an algorithm that inherently avoids one individual receiving an inappropriate number of selections when the sampling rate is low. Essentially, the algorithm only selects a single starting point at random (rather than multiple points) and selects parents each time an integer boundary is crossed.

Turning to the average performance, it was pleasing to see that the average MSE measure followed an asymptotic curve with increasing generation. The derived function provides a method for predicting the appropriate number of evolutionary cycles necessary to reduce most of the average error. In fact, 90% of the reduction in MSE occurred during the first 6 generations. Of course, this figure may only be valid for the current settings. However, it does indicate that convergence is possible and can occur over a relatively few evolutionary cycles. Of course, the slope of MSE curve *could* have been much shallower, meaning that convergence would have taken longer.

One of the design criteria is that convergence on an acceptable face be made as quickly as possible. Firstly, a witness is being presented with multiple faces and this format *could* have an adverse effect on the internal facial representation. The literature at first appears mixed on this issue. For example, Maudlin & Laughery (1981) find facilitation in recognition performance following the construction of an Identikit. No significant effect was observed by Davies et al. (1978) using the Photofit system, but Hall (1977) [cited in Maudlin & Laughery (1981)] found significantly worse recognition performance following the production of a suspect from a sketch artist. An important pattern emerges here: the more realistic the representation being constructed (line drawing elements to photographic elements to sketches), the worse the following recognition performance. If recognition performance can be related to the integrity of the internal facial representation, then caution must be observed in systems that attempt to produce more lifelike photofits (like that from a sketch artist and also the photofit

system under development here). Exposing such high quality facial material to a witness should therefore be minimised. All else kept constant, the number of generations should be limited as far as possible.

In use, photofit systems still arrive at a solution iteratively. There is a clear starting point and changes are made to a single face until an acceptable likeness is achieved. Likewise, in the evolutionary approach, the "end point" would have been reached when one of the population faces is of acceptable quality. Although in use, a witness is likely to point out when a suitable likeness has been reached, the equivalent in Experiment 1 is seen to have occurred when a subject assigns a maximum value on the rating slider; recorded as a rating of 10 points. One can then get an indication of likely system success then by examining the number of maximum scores. In total, maximum rating was found nearly 50% (47%) of the time. If the top quartile ratings can be attributable as a "close enough likeness" (i.e. scores of 8 or more), success can be seen for 90% of the time. Overall, these data strongly suggest that the simple system is performing rather well. It is worth noting too that this does not appear to be a chance result due to the significant negative correlation between the maximum rating scores and MSE.

Turning to the timing data, overall, it was found that it took 5.4 seconds on average to rate each face (32.2/6 seconds). There was no difference in time taken to rate faces in isolation (Condition A) or in the context of the other faces (Condition B). If rating time is linearly scaleable, this means that for much larger populations, rating can be accomplished rapidly. For example, if 18 faces were used, as in Peter Hancock's (2000) prototype system, then rating could be achieved within a couple of minutes (5.4 seconds x 18 faces = 97 seconds = 1.6 minutes).

Summary and Further Work

This chapter has examined a framework for evaluating a holistic photofit approach. It was decided that design should proceed on a P.C. running Windows NT with Microsoft Visual C++ and Matlab as the major software tools. Design of the Pilot System was detailed. Many simplifications were employed to limit the software engineering effort at this early stage. These included building a PCA model from a small database of full-frontal male Caucasian faces (35) of normalized shape (*shape-free*). A simple Genetic Algorithm operated for 12 generations on subject ratings from a small population of faces (i.e. 6).

It was hoped that rating faces in the context of other faces would provide a preferable fitness method for the GA: no differences were found in terms of errors scores (MSE), rating scores and time to rate. Overall, performance was shown to be highly encouraging though, with the average MSE scores reducing asymptotically by a value of approximately 55. Interestingly, subjects were found to assign a maximum rating to at least one of the population

faces in 50% of the targets and a rating of 8 or more for 90% of the targets. In addition, a maximum rating was assigned on average before the 5th evolutionary generation.

These results provide further evidence to allay concerns raised in the previous chapter regarding the ability to make facial similarity judgments. Recall that the evidence presented was based primarily on the observation that confusion between faces was not only common between subjects but also predictable (Courtois & Mueller, 1981; Davies, Shepherd & Ellis, 1979; Goldstein, Stephenson & Chance, 1977; and Laughery, Fessler, Lenorovitz & Yoblick, 1974). Clearly, that the Pilot System was able to evolve faces to become more like a target, not only indicates that subjects are able to make appropriate similarity judgments but also that the system is sufficiently sensitive to capitalize on this information.

There are clearly several areas where improvement could be made to the design. One area is the implementation of gross shape changes. Recall, that faces used in the pilot had all features “fixed” in a pre-specified location; so-called *shape-free* faces. In itself, limiting the configural³² changes in this way may have increased the difficulty of the task. Subjectively, several subjects commented on the apparent similarity of the faces. Obviously, much of the variation between faces is lost when merely the “texture” information is modelled. A better design would model the variation in facial shape, including (a) the shape of facial features and (b) the spatial relationship between facial features. This design is therefore proposed as one of the next developmental improvements.

Another area of improvement is in the collection of user facial fitness information. It would appear that rating scales can be used inappropriately. It was decided that “backwards” rating could be overcome by the use of operator input and scaling techniques could improve the effectiveness of the scale. However, if a system is designed that will be evaluated without the use of constant supervision, a different method of facial selection should be considered. One method would be to use an “anchored” rating scale, containing labels to define the scale. The simplest type perhaps is to label the “end points”; for example one might use “low similarity” for the left-hand end and “high similarity” for the other.

A further problem area was in the use of the Roulette Wheel sampling algorithm, causing an inappropriate number of “poor quality” faces to be selected as parents. This undesirable effect occurs due to the small number of parents that were selected (i.e. 12). One remedy would be to increase the population size, naturally resulting in the selection of more parents. Increasing the number of parents is likely to be a valuable improvement from the

³² Note that this term is ambiguous. Bruce (1995) explains that it can refer to the interaction of features (e.g, the perception of the mouth being altered by the shape of the nose); the “holistic” processing of faces; or the spatial relationships between facial features. It is this last definition, the spatial relationships between features, that is meant when the term “configural” is mentioned in this thesis.

perspective of the GA since this increases the number of points in “face space” and raises the probability of finding an acceptable solution (Goldberg, 1989). Of course, this needs to be balanced against the potential interference effects caused by over-exposure to faces (as mentioned previously).

Improvements can be made in the use of a different sampling algorithm as well. The algorithm suggested by Baker (1987) would appear appropriate for implementation in further design. Note, this algorithm could be implemented along with an increase in population size to ensure that unwanted over-sampling still does not occur.

The next chapter serves to implement and then evaluate these proposed changes.

Chapter 3: Full-Face System (Mark II Face Evolver)

This chapter develops the Face Evolver software to evolve faces that change appropriately in both shape and texture. Evaluation is carried out with a larger population size than before, a simpler method of facial selection and a range of hairstyles. User rating scores indicate significant success over even a few evolutionary cycles (e.g. four generations), though targets obtained external to the face model did not perform as well. It is shown that the poor availability/selection of hair is a likely reason for this decrement in performance; another, is the simplified face model. It was estimated that 10 generations would be required to evolve one of the population faces to the category of "Faces could be easily confused", indicating a likely upper operating limit, though this figure could be reduced by increasing the number of faces in a population. Overall, the approach is once again found to demonstrate considerable promise and serves to promote further development in future chapters.

Increasing Utility

The Pilot System (the Mark I Face Evolver) evaluated in the previous chapter indicated promise for a GA/Holistic approach used as a basis for a new photofit system. This was based on a significant improvement in overall rating and error scores with increasing generation, plus a high proportion of maximum ratings assigned for each target. This system is far from being useful in a forensic setting. Arguably one of the most important developments is a shape model, permitting statistical changes typically found in the relationship between facial features. Simply, a facial shape model is now required in addition to a texture model.

An associated issue is that the faces produced in the previous chapter contained just the eyebrows, eyes, nose and mouth - the so-called "internal features" of a face (Bruce, 1988). Of course, faces in general also contain hair, ears and an outline of the head - the "external features" of a face. This limitation was imposed to simplify the programming effort. A natural effect of adding a shape-variant model and the external facial features is an expansion in the complexity of the face space. This would suggest that a "photofit" would tend to be further away from initial solutions in face space, necessitating longer search times.

This chapter designs a full shape-variant face model (referred to as the Mark II Face Evolver). About this time, opportunities became available to design an exhibit demonstrating the principles of evolution with faces. The exhibit would be resident in a public gallery (at the Hatton Gallery, University of Newcastle) for a total of six weeks and permission was sought to record performance data. This permitted several exhibits to be designed, with evaluation. Part of the design strategy in the following section is based on the notion that such an exhibit would be run by members of the public without supervision. Considerable care is taken to minimize problems experienced previously with unsupervised operation (especially through the use of unanchored rating scales).

Design of a Mark II System

Face Shape Model

A primary improvement to the Pilot System would be the addition of changes to the gross *shape* of generated faces. In Hancock (2000), so-called *shape-free* faces were produced initially followed by the application of a shape morph. The shape morph was derived from a shape model built using PCA of the coordinate locations used to define the major facial landmarks. As with the texture, the *shape* model starts by normalizing the set of image coordinates (i.e. one coordinate vector per face) to zero mean and unity standard deviation. This is followed by the computation of the covariance matrix and the extraction of the eigenvectors. *Eigenshapes* (cf. *eigenfaces*) are computed by multiplying the eigenvectors by the original coordinate sets. The *eigenshapes* therefore capture the gross shape changes in the database of faces.

Much as before with the texture model, construction of a novel face shape proceeds by the weighted addition of the eigenfaces to the average shaped coordinate vector. The weightings are produced from a vector drawn from a Gaussian random number generator with each element scaled to its corresponding eigenvalue. Production of a novel *shape-free* face (from the texture model) is followed by a bilinear interpolated shape distortion (i.e. a *morph*) from the average shape to that specified by the novel coordinate vector.

Hair

An inherent difficulty still remains with this approach regarding hair. The problem is that although facial shape and texture information tend to be largely consistent over time, more so for men than women because of the general reluctance of men to use makeup in this country, this is clearly not the case with respect to a person's hair. It was considered best therefore that hair should be considered an independent feature to the face because the colour, length and style can be easily changed. A solution is proposed such that the structure and intensity of a face be modelled holistically, but the hair be represented in a more feature-based way.

Clearly then, a diverse range of hairstyles should be available to accommodate the needs of a witness. A problem is how to include a selection of hairstyles that form part of the external features and can be fused with the internal features to produce a single face. Arguably one could assemble a large repertoire of hairstyles (say from photographs), though the easiest approach initially might be to make use of the hairstyles that form the corpus of faces. At this stage in development, there were 35 faces in the corpus and these hairstyles are all quite short and tidy. Naturally, this is quite a restricted set but can be of value at this early stage of system development.

Targets

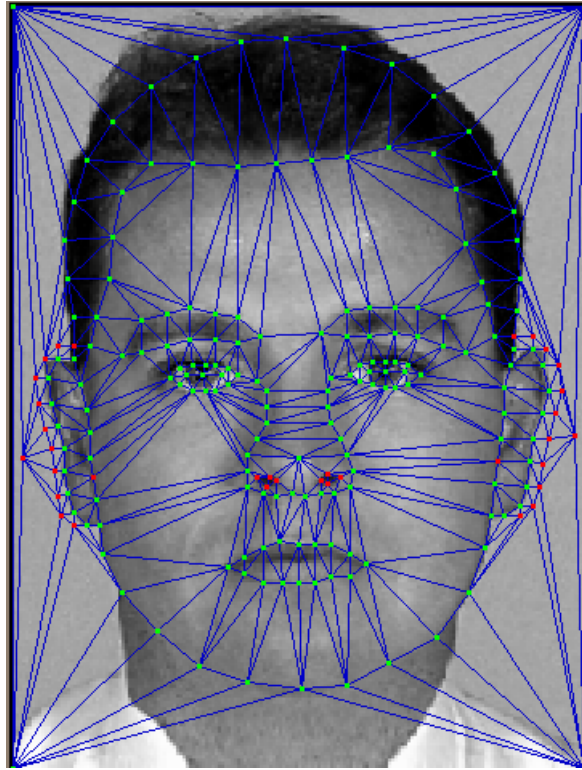
The design of the Mark I Pilot System considered the origin of the target images. Having the target generated from within the system explored whether it was possible to locate a face in the texture model. This sensible approach is continued to include both the shape model and also hairstyles as well. Once again, a face could be randomly generated and one of the hairstyles selected (at random) from the database as well. In fact, the entire set of external features from the reference images could be used, providing not just the hair but also a neck, collar-line and a pair of ears.

Recall that the original set of images were first shape-normalized (to the average shape of the corpus) prior to creation of the texture model. To create a novel face, a randomly generated *shape-free* face would first be generated (as in the Pilot System) and its internal features then inserted into the selected shape-normalized reference background. The final step would be to apply a morph to this face with coordinates specified by a random location in the shape model. The result would be a randomly generated full shape-variant face with a hairstyle chosen from one of the corpus images. Like the Pilot System initially, both the population faces and target face could be created using this method (i.e. internal to the system).

Improving Image Quality

When this was implemented however, several difficulties were observed. The first concerns the number of coordinates used to mark key facial locations. Only forty-two were used in the Pilot System and these were positioned mainly around the eyes, eyebrows, nose and mouth. This was a reasonable number considering that only the internal features were being displayed. The number of points marked on the external features was limited with the result that the outline of the head appeared jagged rather than a smooth contour. To model the external features more acceptably, the number of coordinates was increased to 211. An example can be seen in Figure 17.

Figure 17: Coordinate Point Locations for the Full Shape-Variant Model



The other problem was that when inserting the randomly generate internal features into the shape-normalized head, an obvious discontinuity was present. A solution to this was found by “blending” the internal and external features at the edges where the internal and external feature overlapped. This was easily achieved by providing a “keying” mask that provided a graded blend across the contour of the internal and external boundary. The following mask was designed –

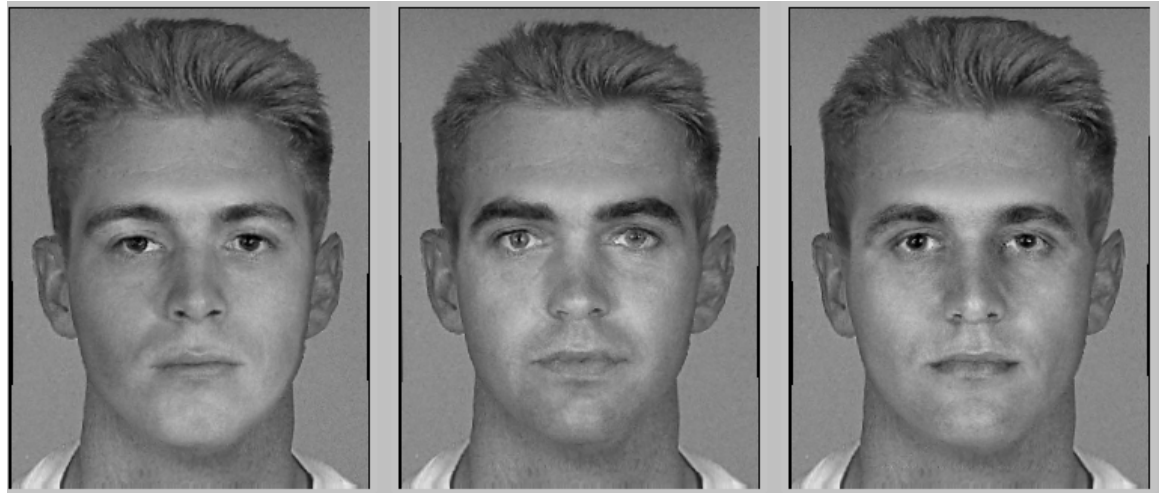
Figure 18: Keying mask used to create a composite image from the internal and external features



This (Figure 18) was created manually in Adobe Photoshop. A composite image is formed exclusively from the external image when the pixels are black in the keying mask, and from the internal image when the pixels are white in the keying mask. A mixture of internal and external feature pixels forms at the boundary (the blurred areas of the mask appearing in grey). To produce good results, it was found that pixel blending needed to occur over considerably more pixels in the forehead and chin areas than in the region between the cheek and ears.

Figure 19 illustrates the effect of keying 3 randomly generated textures into one of the shape-normalized backgrounds and then applying a randomly generated face shape. Note that the blending is very acceptable and the resulting faces are of high quality.

Figure 19: Examples of Randomly Generated Shape-Variant Faces



Demo System

It was mentioned earlier that there was an opportunity at this point in the project to provide a public demonstration of how an evolutionary process could be applied to a small population of faces. Since supervision could not be guaranteed at all times, the exhibit would need to be self running and sufficiently engaging for members of the public (who would generally be unaware of the project). In addition, the exhibit should take only a few minutes to complete (otherwise users are likely to get bored) and feedback should be given immediately after the evolution process. Permission was obtained for information to be collected regarding evolutionary performance, thereby enabling some evaluation of the new system to be carried out.

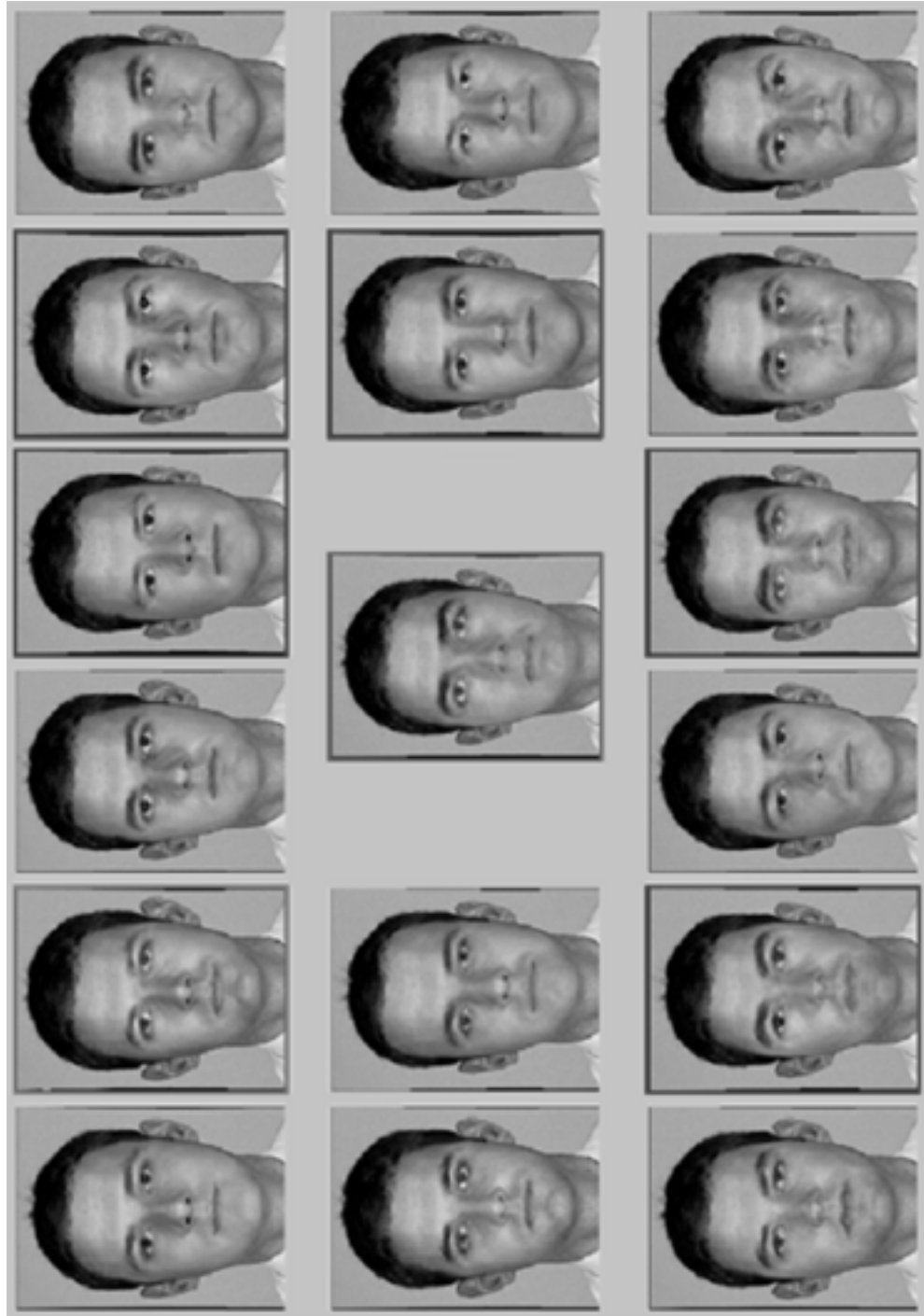
The philosophy adopted in the Pilot Study, with the target present during evolution, was thought appropriate. Once again this would serve to explore whether the system was able to locate targets in the shape and texture models. In addition, the task of evolving from memory may be too hard and therefore not appropriate in an exhibition setting.

One design factor thought sensible was to increase the number of faces in the population. Previously there were 6 faces, displayed as two rows of three (Condition B). It would be preferable to increase the population size as much as possible (so as to increase the number of solutions being explored in parallel by the GA). To avoid the potential confusion of seeing too many faces across too many screens, the population size was limited to that which could fit on single screen. The maximum number of faces displayed is of course limited by the size of the monitor: the more images displayed on a screen, the smaller they will tend to be. Ultimately, it was thought that images seen at least 45mm by 60mm on the monitor (i.e. *physical image size*) would allow general use of the exhibit.

The highest density turned out to be a configuration with faces in three rows. As a 17" monitor was to be used, this enabled a maximum of 3 rows of 6 faces to be displayed within

the necessary size constraints – the same as Hancock (2000). The utility of this layout is underscored by Baker & Seltzer's (1994) successful line drawing evolver and Baker & Seltzer's (1998) successful mugshot album search that both employed a similar number of examples (20) on the same sized monitor. However, it was required that the target be presented along with the population faces. It seemed most natural to position the target in the middle of the display and have the population in the surrounding area. As there was not an odd number of faces in a row, the middle two population faces in the second row were replaced by the target. This resulted in sixteen population faces. An example illustrating this configuration is shown in Figure 20 –

Figure 20: Presentation Format for Faces (the target is displayed in the centre of the population)



With this interface designed, concern was expressed as to the method by which users should weight the population fitness function. Recall that in the previous chapter, this was achieved via a simple 10-point rating scale. However, the lack of constant supervision suggests that this might once again result in undesirable rating behaviours. It was thought that a very simple method would be to ask people to merely select faces that they thought were closest to the target – rather than specify a rating.

This approach is rather like the apparently successful method adopted by Rakover & Cahlon (1989), though more faces could be selected in the current paradigm. The whole face selection method has featured in several other studies demonstrating utility in searching for targets (from memory) in mugshot albums (Baker & Seltzer 1998; and Levi, Jungman, Ginton, Aperman & Noble, 1995). In Levi et al's (1995) feature-based retrieval system, a database of 1200 faces was searched for a target by displaying a screen of 24 faces and having the witness select the closest five. These selections were used to adjust the weightings of the features to produce a further set of faces for selection. The process would continue until a match was achieved (occurring at least twice as fast as a traditional, linear search and with an 80% success rate). Baker & Seltzer (1998) also required five selections, this time from a set of 100 randomly chosen mugshots. Interestingly, the authors performed PCA on this 4500 item photo album and located the position of the target in this sub-space closest to the highest ranked face. It was found that the target occurred on average 3 times sooner compared with a linear search and with about an 80% success rate. The authors go on demonstrate even better performance (a higher hit rate and lower search time) if a top "composite" is used as a reference (instead of the highest ranked face). This image is selected as the best from 10 composites assembled randomly from the facial features of the five best selected mugshots.

Overall, Baker & Seltzer (1998), Levi, et al. (1995) and Rakover & Cahlon (1989) illustrate the utility of making similarity judgments to a target. The implication is that a whole face selection mechanism may well be valuable for the Face Evolver. Further, that the selection mechanism has been used in a PCA setting, admittedly for face recognition rather than generation, provides additional confidence.

Returning to the Face Evolver's interface, this mechanism could be easily carried out by simply clicking on a face with the computer's mouse. Selected faces could be indicated by changing the image's border, providing sensible user feedback. A second click on the face could de-select it; useful when a mistake had been made or when a "change of mind" was required.

Proportional fitness selection could be used again, with each selected face receiving an equal opportunity to take part in the breeding process. It was perceived however that some people might not select many faces at all – just selecting only one or two. The likely result would be that the population would converge early and not necessarily on the given target. A minimum should therefore be imposed on the number of faces selected. This minimum should be large enough to guard against early convergence but not so large so as to take too much time and be boring for a participant. After an early trial version, it emerged that people were generally comfortable with selecting half a dozen (six) faces. Although this requirement would be stipulated in the instructions, it was decided that user feedback would be provided when less than six faces had been chosen (by way of a message box that requested the selection of more faces).

An observation made from the Pilot System was that about half the participants gave a maximum rating sometime during evolution. This suggests that it might be appropriate to indicate those faces thought to be very good. One method would be to allow participants to make an initial selection of the face that was considered “perceptually closest” to the given target; Goldberg (1989) refers to such individuals as “best-of-generation”. This could then be followed by the selection of the remaining five. This simple approach results in a minimum of 6 mouse clicks per generation. Selecting the “best” face at the start is an ergonomic solution since it allows more than 6 faces to be easily selected should a subject desire. Mechanisms should be put into place that would permit the simple re-selection of the best face should a change be necessary. Arguably, the simplest way would be to click on the best face again, like the de-selection of a population face, followed by a click on another face, assigning that as the best.

The selection of a best face also permits another mechanism in GAs to be used: elitism. In elitism, one carries forward to the next generation the individual(s) that were considered superior. This based on the notion that particularly good individuals are likely to be beneficial in future generations, since they have been of relative value previously, and also to avoid the population from becoming qualitatively worse (since the best face from a generation would be, at worst, still the best face in the next generation). So, it was decided to include the best face in the following generation. Therefore, the GA would operate by the *replacement* of fifteen parents from a previous generation, leaving the sixteenth as the best face from that last generation. Finally, the best face would be positioned in a random position within the population so as to remove any positional cues by placing this face in a fixed location.

In addition to elitism, the best face could also be given a greater weighting than the other faces selected. In GA terms, the “selection pressure” would then be higher for this face, resulting in more breeding opportunities. This would lead to more offspring being produced with the influence from this “preferential” individual. The consequence is a decrease in convergence time. Although the amount of selection pressure is unknown at present, it is proposed simply that twice the weighting be attributed to the best face (i.e. a 2:1 selection pressure).

The presence of this best face would allow another method of system evaluation. This could involve asking participants to rate their best face against the target face. The target face could be present during the rating exercise to avoid confounding effects by “holding” the target in memory. Once again, one would want to avoid the inappropriate use of rating scales. Therefore an “anchored” scale was suggested with a set of categories ranging conceptually from “a very poor similarity” to a “perfect match” with the target.

It was important that this scale should be easy to use as members of the public would be the subjects. The most appealing design was a fully anchored scale, to encourage consistency between subject. Ultimately, a scale was designed containing 6 major categories and a single division within each of the intermediate categories for finer discrimination. A

small pilot study indicated that members of the public would be able to use it with ease. This is referred to hereafter as the Anchored Face Similarity Scale or *AFSS* and is shown in Table 2 -

1	Very poor likeness between faces
2 or 3	Few similarities
4 or 5	Some similarities
6 or 7	Many similarities
8 or 9	Faces could be easily confused
10	Faces are identical

Table 2: The Anchored Face Similarity Scale (AFSS) used to Evaluate Performance

The rating procedure has a further advantage of giving the user something to do while the faces are being computed, which took 4.5s on average³³ using the P.C. that ran the Pilot System (refer to the Apparatus section, Chapter 2).

Clearly, a good exhibit would be one that was able to demonstrate a desired effect in a short time. In this case, one would like to be able to show that the evolutionary system had become more similar to the presented target over a few evolutionary cycles. In the Pilot Study, peak performance was observed after 4 generations (as measured by the average generation that a maximum rating was assigned). Using this result as a guide, it was proposed to run participants through 4 generations of the software initially. However, to allow opportunities for further evolution for participants that were prepared to persevere, an option was proposed that enabled continuation of another 4 cycles. This has the added advantage of potentially being able to gather data for longer runs of the system.

It is of importance to demonstrate evolution of a face for the exhibit. After 4 or 8 cycles of the software, several measures could be presented to the user to demonstrate evolutionary success. These could include reporting the subject's rating scores or even the error score (MSE with respect to the target) of the best faces. Even the groups of selected faces could be shown, as these are likely to have changed with increasing generation. But, these measures are either too abstract (as with rating or MSE scores) or too confusing (as in the case of showing groups of selected faces). Arguably, the simplest method would be to show the progression of best faces over time along with the target face (to permit direct comparison). Of course if the system is working, then the best face should become perceptually more similar to the target.

It was expected that some subjects would not continue evolving to 4 generations. This could be on account that they became bored, had to leave the exhibition, or most importantly, they were not serious about the exhibit. It was felt members of the public who were not serious about the exhibit are unlikely to provide useful insights into the Face Evolver's performance. To guard against this, it was decided to collect and analyze data only from those subjects who completed the minimum number of evolutions. It was also decided that participant

³³ This was based on the average (mean) of 6 replications of the evolutionary generator.

demographic information should be collected, though this should be optional and minimal: age, gender and user comments. It appeared best to collect this information following the presentation of the participant's best faces. Opportunity for user feedback was considered a good idea, due to the lack of continual supervision.

Experiment 2: Shape-Variant Face Model

This experiment provides the first evaluation of the Mark II Face Evolver. It differs from the Mark I version by the inclusion of major facial shape variations, the addition of hair, user selection by faces (as opposed to rating scales), a larger population size (now 16 faces), and the use of Baker's algorithm (to avoid inappropriate over-sampling of low rated faces). The evaluation is to be conducted in a public setting with targets generated internally and evolution continued over 4 or 8 generations. Assessment of performance is planned via the use of an anchored rating scale of the most fit (or best) individuals from each population.

Method

Overall, the Mark II Face Evolver was designed to parallel the Pilot System as much as possible - to facilitate comparison between studies as far as possible³⁴. Hence, the PCA *shape-free* (texture) model from the Pilot Study was carried forward and the first 17 coefficients (plus the average database image) were used for image generation.

A PCA *shape* model was built from the original set of image coordinates and the first 17 of these coefficients were used for shape generation (i.e. the same number as for texture generation). Creation of the 16 random faces and the target face proceeded with the *shape-free* images produced as before followed by the keying of the internal features into one of the 35 original *shape-free* external features using the blending mask described previously (refer to Figure 18). A final shape de-normalizing morph was carried out by a shape vector defined by populating the first 17 coefficients with random numbers (drawn as normal from a uniform Gaussian distribution and scaled by the corresponding eigenvalues) from the shape model. The face selection procedure adopted, in contrast to rating scales used in Experiment 1, allowed each face to have an equal share of being a parent; the exception being the first (or best) selected face - receiving a weighting of twice the others (a higher selection pressure). Proportional selection fitness and uniform cross-over for both the shape and texture components was once again adopted, though Baker's algorithm was implemented in place of a Roulette Wheel method of parent selection; mutation was set to a probability of 0.05, replacing the shape and texture coefficients with an appropriately scaled random value.

³⁴ Although the creation of full shape-variant faces, a different rating method and Baker's algorithm (for parent selection) may serve to cloud comparison.

Initial analysis of the system was planned using the rating scores of the best face assigned at the end of each evolutionary generation. Of particular interest is the distribution of “high” rating scores and whether significant differences could be found with increasing generation. The presence of significant effects would be of interest to the eighth generation as well, should sufficient participants decide to continue that far.

Participants

Twenty-two members of the public participated in the study, set up as an exhibit during the Science of the Face exhibition at the Hatton Gallery, Newcastle University (June-July 1999). Participation was voluntary. There were 13 males and 9 females. Their ages ranged from 10 to 43. 10 participants continued evolution beyond the fourth generation (the remainder terminated the evolution process after 4 generations).

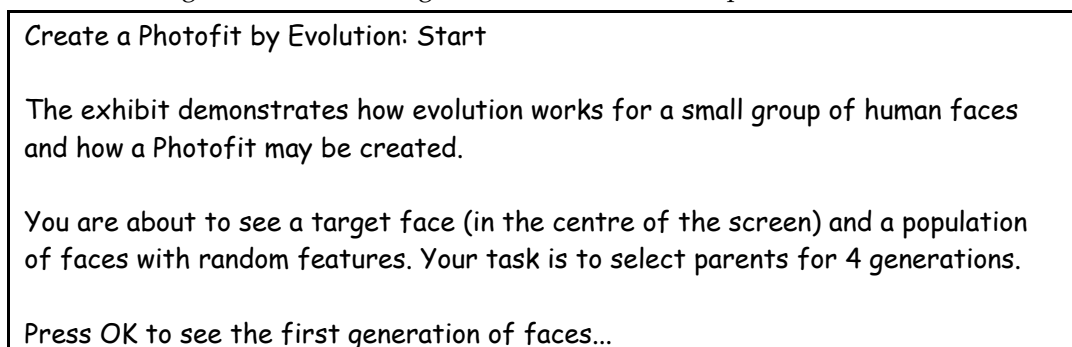
Apparatus

A Pentium PII PC clocked at 350MHz was used to run the experiment. Faces were displayed on an Iiyama 17” monitor. Participants had use of the computer mouse (to select faces and move to the next screen) and keyboard (to enter rating scores and complete a short demographics section at the end of the experiment).

Procedure

Each participant was assigned to the next random target (a different set of external features were used for each subject, chosen sequentially from the original shape-free corpus). The first generation of faces presented to the participant was generated randomly. Instructions were presented at the start of each trial (via a Window’s Message box) that explained to first select the face that most closely resembled the target face, then 5 others that closely matched (Figure 21) -

Figure 21: User Message Presented at Start of Experiment



Closing of this message displayed the population faces accompanied by the following message box to reinforce the required task (Figure 22) -

Figure 22: User Message Presented In Front of the First Set of Faces

These are an initial set faces with random features.

What do I do now?

1. Look at the target in the centre of the screen (behind these instructions)
2. First, click on the face *MOST* similar to the target
3. Then, click on 5 other faces (with similar features to the target)
4. Finally, click the *Make Next Generation* button

While the next generation of faces was being computed, the target was shown together with the user's closest selected face (best face) and rating for similarity was carried out (using the AFSS, Table 2). To further reinforce the task, the following message was displayed in front of the second generation of faces (Figure 23) -

Figure 23: User Message Presented In Front of the Second Set of Faces

These are the 'offspring' faces

What do I do now?

1. As before, first select the face *MOST* similar to the target, then 5 others
2. Click the *Make Next Generation* button when done.

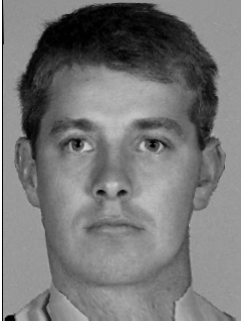

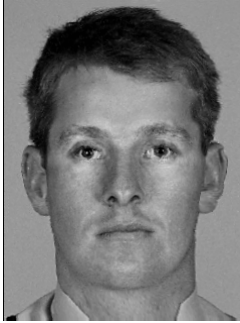

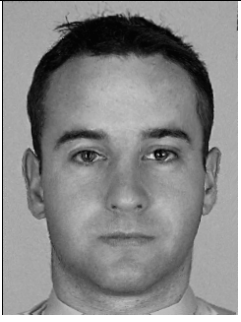
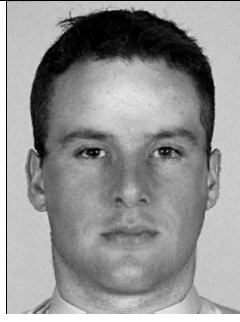



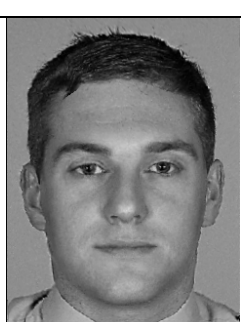
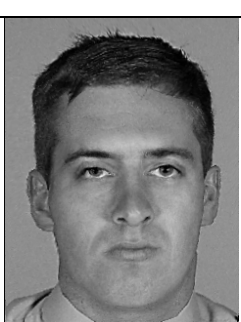

When 4 generations of faces had been evolved (the initial set plus three more), a request was made to continue for another 4 generations. When the evolution was complete, the set of best faces was displayed on a screen along with the target face. Demographic information (age and gender) was then collected along with user feedback. Debriefing involved explanation of the evolutionary process and its application in a forensic setting. The rating data and all faces were saved for participants that completed 4 or 8 generations; data from incomplete trials were over-written (with the next participant).

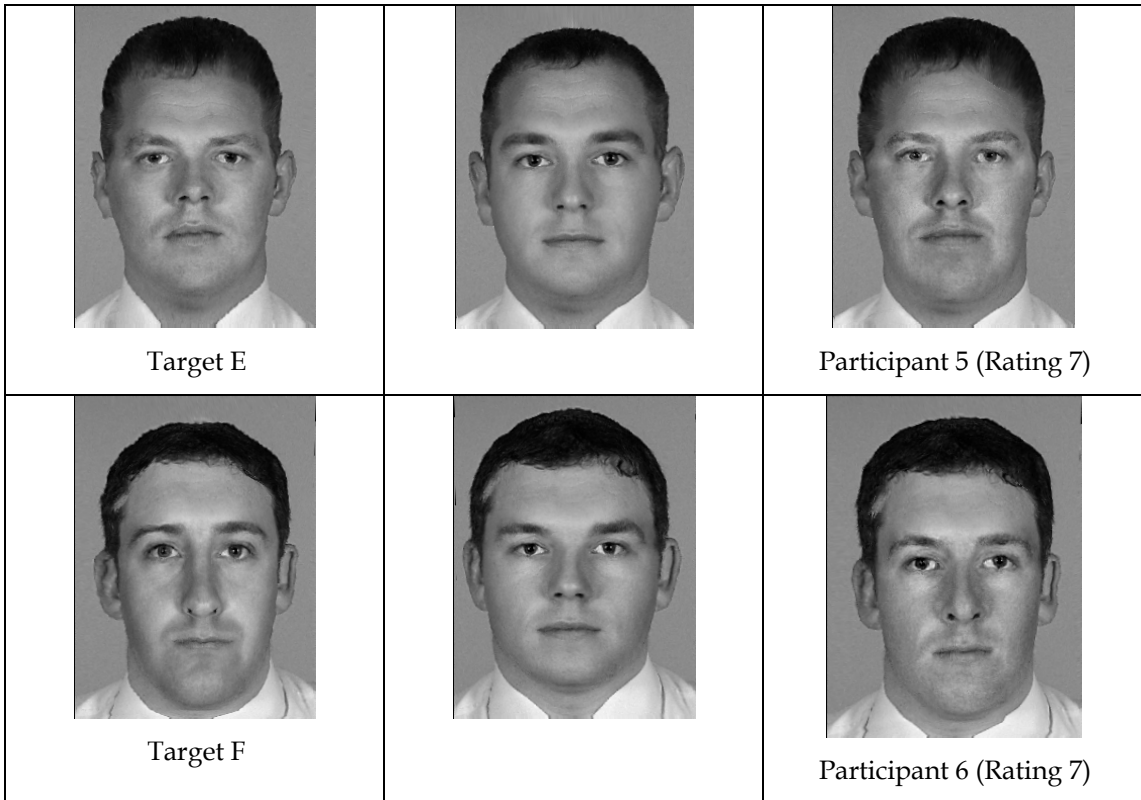
Results

Figure 24 shows a few examples of performance from the first six participants. A significant improvement can be seen subjectively from a "typical" face³⁵ in the first generation to the best-rated face in a later generation -

³⁵ This was an example (other than the best face) taken from the first generation.

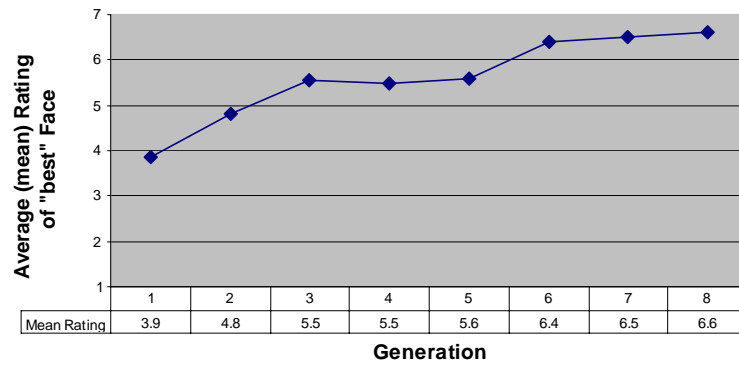
Figure 24: Examples of Evolutionary Performance

Target	Generation 1: Typical Face ³⁵	Best rated face
 <p data-bbox="416 622 529 656">Target A</p>		 <p data-bbox="1098 622 1370 656">Participant 1 (Rating 9)</p>
 <p data-bbox="416 1003 529 1037">Target B</p>		 <p data-bbox="1098 1003 1370 1037">Participant 2 (Rating 8)</p>
 <p data-bbox="416 1400 529 1433">Target C</p>		 <p data-bbox="1098 1400 1370 1433">Participant 3 (Rating 9)</p>
 <p data-bbox="416 1785 529 1818">Target D</p>		 <p data-bbox="1098 1785 1370 1818">Participant 4 (Rating 9)</p>



The mean rating scores for the best faces have been plotted in Figure 25 below -

Figure 25: Improvement in Rating Scores

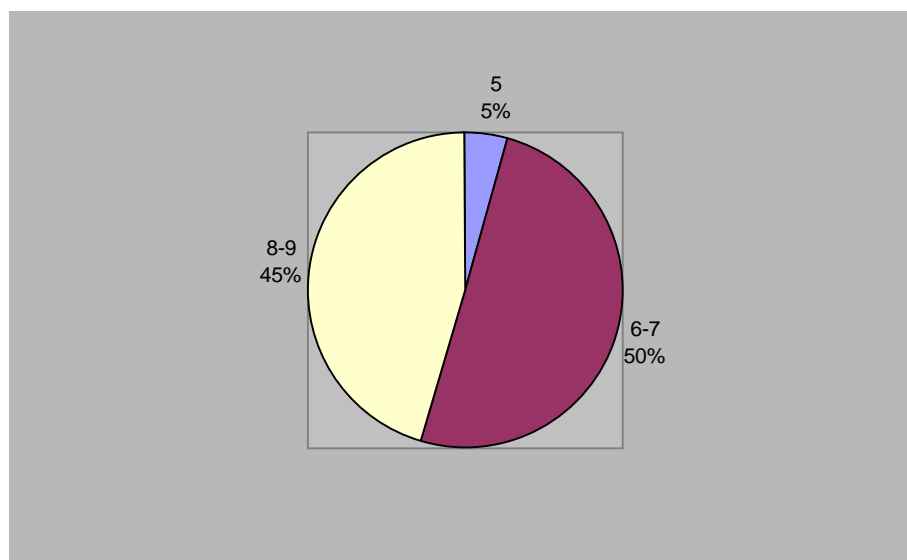


Looking at the change in mean rating scores with successive generations, scores increased for the first two generations, remained nominally constant for a further two generations, increased and then remained roughly constant for the remaining time. Mean rating scores increased from 3.9 in the first generation to 5.5 in the fourth generation to 6.6 in

the eighth generation. A repeated-measures ANOVA for rating scores³⁶ found a significant main effect of generation collected over the first 4 generations ($F=11.4$, $DF=(3,21)$, $p=0.003$). A two-tailed within-subjects t-test³⁷ revealed that rating scores in generation 4 were significantly higher than generation 1 ($t=3.40$, $DF=21$, $p=0.003$); for simplicity, no other t-tests were performed. A repeated-measures ANOVA also found a significant main effect of generation for rating scores between generations 5 and 8 ($F=18.5$, $DF=(3,9)$, $p=0.002$) and a t-test indicated that rating scores in generation 8 were significantly higher than in generation 4 ($t=2.8$, $DF=9$, $p=0.022$).

Consider the distribution of the highest score attributed to a best face for each participant. It can be seen from Figure 26 that approximately half the time the best scores occurred for rating of 6-7 ("many similarities") and half the time for ratings of 8-9 ("faces could be easily confused"); only one subject rated below either of these (in the upper category for "some similarities"). There were no maximum ratings assigned.

Figure 26: Distribution of the highest rating for each participant



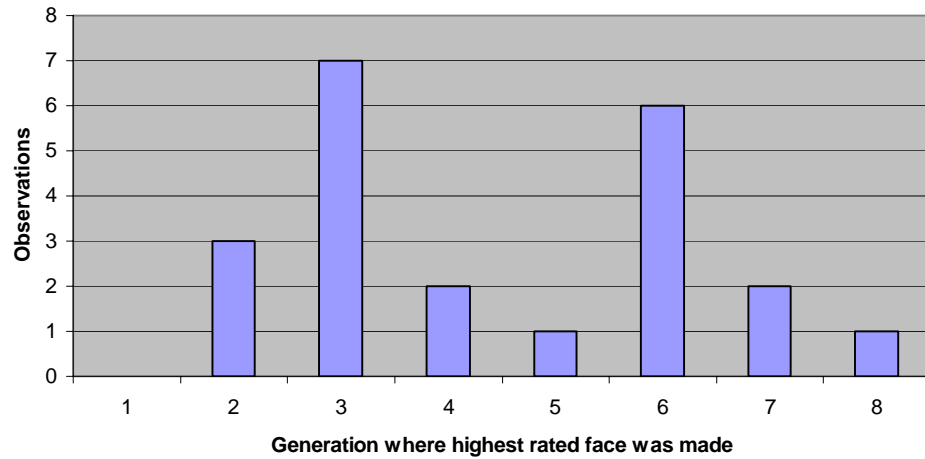
Consider next where these highest ratings were made. As it was sometimes the case that there was more than one face with an equally high rating, analysis considered the face

³⁶ Although it is acknowledged that non-parametric statistics are preferred with rating scores (as the data is only ordinal and not (at least) interval), it is common practice in the psychological literature for parametric statistics to be used. Following this approach, parametric tests will be carried out for rating data throughout this thesis.

³⁷ For ease of readability, all subsequent t-tests in *this* chapter are the "two-tailed within-subjects" type unless otherwise specified.

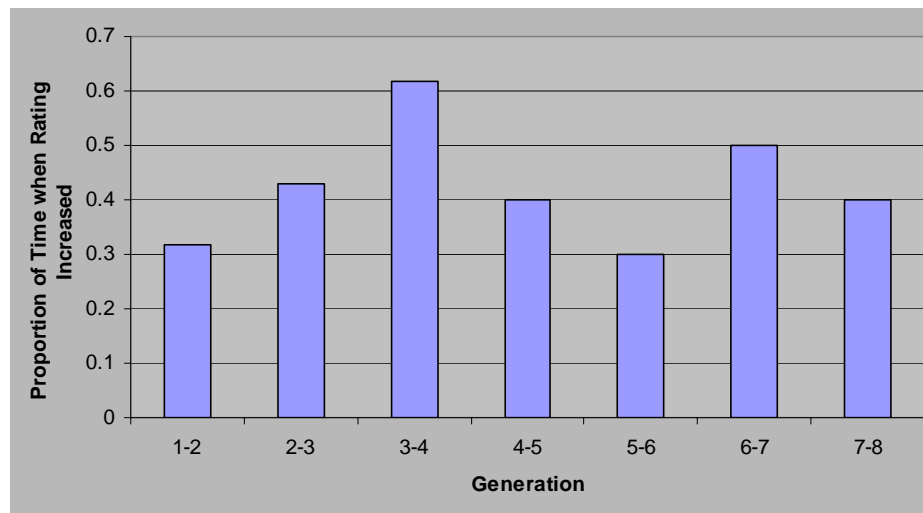
from the higher generation³⁸. The distribution where the highest rated face occurred reveals two clear peaks (Figure 27) occurring in generations 3 and 6 -

Figure 27: Distribution of the Highest Rated Faces



In addition, there were 79 out of 106 occasions (74.5%) when participant ratings either increased or remained constant from one generation to the next; 42.4% of the total time scores increased and 32.1% it remained constant. Figure 28 indicates that increases in rating scores occurred proportionally more of the time especially from the third to the fourth generations -

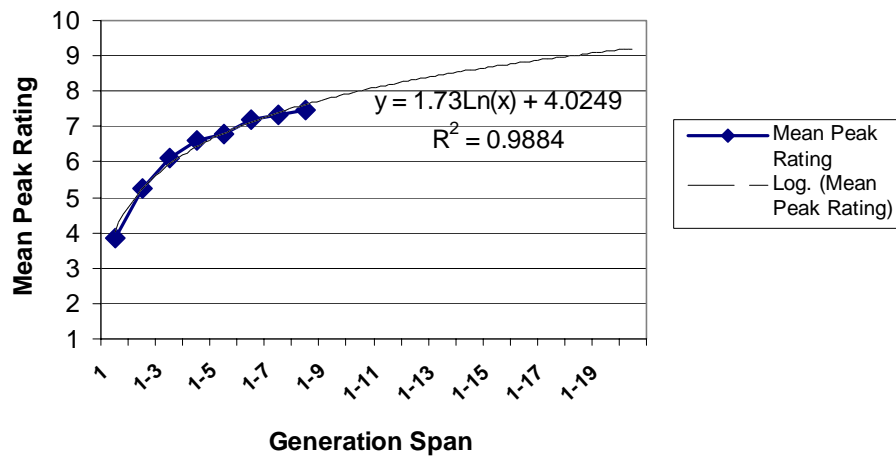
Figure 28: Proportion of Time that Participant Ratings Increased



³⁸ This was based on the assumption that, as system performance tends to increase with generation, a preferable face is likely to occur in a later generation. That said, one reason for an equally high rating is that the same best face is re-selected in the following generation (recall that the elitist mechanism includes a best face from a previous generation) and the same rating applied. In fact, this was found to occur 16% of the time; 2.6 times greater than chance (equal to 1/16 or 6.25%).

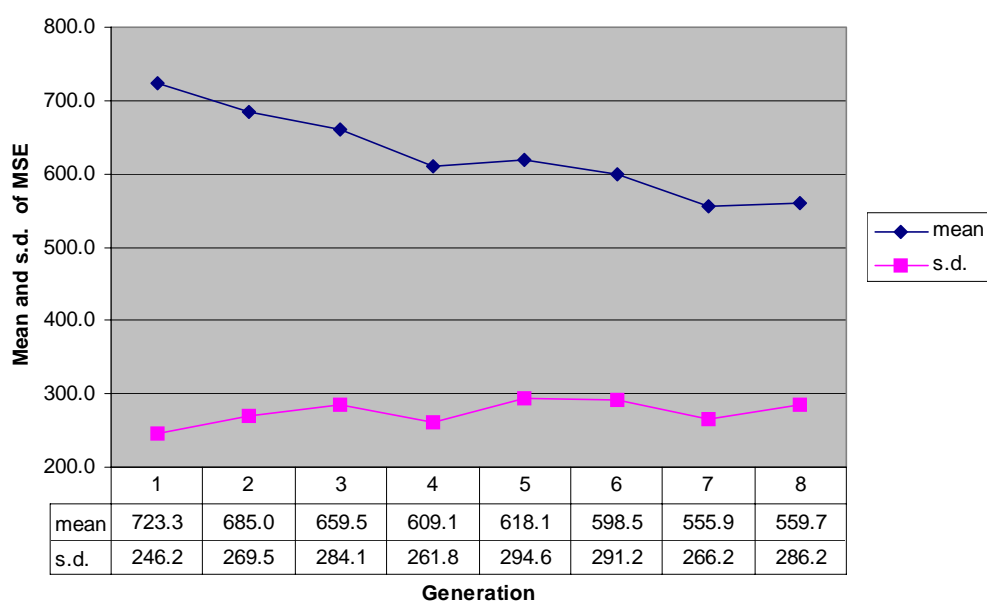
The increase in peak rating scores was then computed over successively longer generation spans. It can be seen from Figure 29 that the average peak rating always increased, although progressively less with increasing generation. A good fit was found for this curve with a logarithmic function (refer to the graph) that explained 99% of the variance in the average peak ratings.

Figure 29: Increase in Peak Rating Scores for Increasingly Longer Generations



Considering the pixel error scores (with respect to the given target), the average and standard deviation of MSE score were computed for each generation for subjects that continued to generation 8. Figure 30 shows that the average MSE scores follow a general trend of decreasing error with increasing generation but the SD scores fluctuate about a mean of about 250 –

Figure 30: Variability in Mean and Standard Deviation of MSE during Evolution



A repeated-measures ANOVA for the MSE data in generations 1 to 8 was found to be significant ($F=11.11$, $DF=(7,9)$, $p=0.009$). Using a t-test, there was a significant decrease between generation 1 (mean 723.33) and 8 (mean 559.72; $t=3.23$, $DF=9$, $p=0.010$); no other tests were computed. An f-test³⁹ indicates that there was no difference in the variance between the first and fourth generations ($DF=21$, $f=1.46$, $p=0.195$).

Discussion

The results of the mean rating scores of the best faces were seen to follow an increasing trend. This was supported by a significant main effect of generation and significant differences were confirmed between generations 1 to 4 and from 4 to 8. This suggests that the best face became perceptually more like the presented target over time and implies that the current evolutionary system is working. The overall size of the effect was found to be a modest 2.7 (out of 10 on the similarity rating scale) from the first to the eighth generation. Relating this to the semantic labels assigned to the rating scale, the average rating increased from the category of “Few similarities” in the first generation to “Many similarities” in the eighth. This provides a measure of the average peak system performance.

Important findings also arise from the analysis of the highest rating score attributed by each participant (i.e. during each evolutionary run of the system). This is relevant, as was argued for the Pilot System, since a high rating can signify convergence to a target and completion of the evolutionary exercise. It was found that maximum best rating for each subject fell roughly equally between the categories of “Many similarities” and “Faces could be

³⁹ An f-test indicates whether there is a significant difference in the variance of two populations.

easily confused". This result is encouraging since it suggests that the target face could be evolved using the system to a high degree of likeness, especially for half the subjects who rated in the "Faces could be easily confused" category. Interestingly, the best-rated face tended to be assigned on the third or the sixth generation. If it can be assumed that the concentration of best-rated faces is indicative of a significant evolutionary jump, then this data fits nicely with Darwin's notion that evolution is slow and gradual with occasional significant adaptations in between (rather than being continuously rapidly increasing); (Dawkins, 1991). This notion that evolution is "slow" was reinforced by the observation that about a third of the time (32%) rating scores did not change from one generation to the next. Impressively, this figure was superseded only by the proportion of time (42%) that rating scores actually *increased*. It was found instructive to compute the increase in peak ratings over successively longer generations. This illustrated the continued contribution of evolution for longer periods of time. A log function was able to fit the curve exceedingly well. A prediction can be made such that 10 generations are sufficient for subjects to rate one of the faces in a population in the category of "Faces could be easily confused." This suggests that 10 is an average operating upper limit for the number of generations required to obtain a high similarity between a target and a photofit.

Further support of evolutionary success is provided by the MSE data. The analyses were computed from the average MSE score from each generation and therefore provide a guide to average population fitness (as opposed to peak performance - as measured by the rating scores of the best face). Clearly the trend was for faces to become significantly closer to the target (in terms of image intensity). This suggests that it is not just the best individual in the population that has improved, but the population itself has become more like the target. For evolutionary success, both of these effects are expected and are present. It was not surprising to find a lack of a significant difference in the standard deviation of the MSE scores. Though a significant difference had been found in Chapter 2, this was over 12 generations and with a much smaller population size (6 faces). The lack of any significant changes in the standard deviation scores in this study suggests that the variation in the population had not changed and implies that continued evolution (beyond 4 or 8 generations) would not be hampered by a lack of population diversity.

Caution should be applied to the interpretation of the rating data however. It is possible for subjects to be driven by experimental expectation. In this case, they could have attributed higher rating scores due to the desire to see "improvement" (or were perhaps being more generous as the demonstration continued). Such behaviours would certainly weaken the above argument suggesting encouraging results. A counter argument to this could be the presence of the two "flattish" regions, or *plateaux*, where a lack of a significant difference was found (i.e. between generations 3-5 and 6-8); the so-called evolutionary "slow" periods. It would appear reasonable to assume that if participants were driven by expectation, then rating scores should *always* rise. It is interesting to note anyway that about a quarter of the time

(25.5%), rating scores decreased and this occurred frequently when participants would terminate the experiment; this occurred 12 out of a total of 27 times (44%) from generation 3 to 4 and from generation 7 to 8. This notion will be further explored in the following experiment.

Experiment 3: Confirmation of Rating Scores

The previous experiment brought into question the validity of user rating scales for evaluation of the Face Evolver. Two effects were presented suggesting that this might not be the case: the presence of two plateaux in the rating distribution and a decrease in rating scores on the final generation in a large number of subjects (44%). The current study examines the ratings obtained from an independent group of subjects.

Method

It was decided that the easiest way to ascertain the validity of the rating scores from Experiment 2 would be to give the best faces to a group of independent subjects to rate for similarity against a target face using the same rating scale. Rating could be carried out with faces presented in a randomized order rather than a serial order (as was the case in Experiment 2). Any differences in rating scores now could then be more confidently attributable to actual differences in system performance as opposed to undesirable subject effects.

As a total of 128 rating scores had been collected⁴⁰, and a small pilot was proposed, only the best faces from the first and last generation of the first 12 participants from Experiment 2 were used as stimuli. The same rating scale would be used to permit valid comparison between studies.

Participants

Sixteen participants agreed to complete the experiment, comprising visitors to the MacRobert Centre at the University of Stirling. As there was hostility regarding the collection of age, no demographic information is available. Participation was voluntary.

Procedure

Each of the 24 best faces (from the first and last generation from the first twelve subjects) was printed on a separate page along with the corresponding target face. These were shuffled for each participant and shown one at a time for rating using the AFSS as before.

⁴⁰ This comprised of four ratings for each of the 22 subjects (for the first 4 generations) and a further four ratings for the ten subjects that continued to the eighth generation ($4 \times 22 + 10 \times 4 = 128$).

Results

The mean rating was 5.26 (SD 2.49) for faces taken from the first generation and 5.76 (SD 2.45) for faces taken from the last generation; this was a significant increase using a t-test ($t=1.97$, $DF=191$, $p=0.005$). Although the average rating scores were significantly higher for the initial generation in Experiment 3 compared with Experiment 2 ($t=3.14$, $DF=214$, $p=0.002$), there was not a significant difference between the rating scores of the final generation ($t=0.68$, $DF=214$, $p=0.496$); between-subjects t-tests were used.

Discussion

The increase in rating for faces between first and final generations provide evidence that the increases reported in Experiment 2 were not solely due to participant expectancy effects. That the effect size was smaller (by 0.33⁴¹), coupled with a significant difference between the rating scores in the initial generation of faces, requires explanation.

The reason for the significantly lower rating scores for faces in Experiment 2 compared with Experiment 3 is likely to be based on methodological differences between the two studies (rather than differences in subject performance caused by the use of different population types – i.e. university students versus members of the public). One obvious difference is that participants in Experiment 2 had exposure to a set of faces *prior* to rating; i.e. they viewed and selected six faces from the initial set of faces prior to rating of the best face. When making a judgment regarding similarity, this is likely to be affected by the variation in the population under comparison. For example, two faces could be considered to be less similar if they were known to be drawn from a family photo album rather than from photographs in general (since smaller differences between family members would be considered more salient and spaced further apart on a rating scale). This in general would allow Experiment 2 participants to gauge the likely variation before making a rating judgment and is likely to be more marked the more dissimilar the faces: especially the first generation. As the presented faces in Experiment 2 represent a sub-set of Caucasian males, and are more similar to each other than faces in general, it is not unreasonable then for Experiment 2 subjects to rate faces in the first generation significantly less than those from Experiment 3.

In conclusion, the presence of a significant increase in rating scores in Experiment 3 does indicate a lack of participant expectancy effects in Experiment 2. The use of rating scales in the manner prescribed (as in Experiment 2) is therefore taken as a valid method of system assessment and that the encouraging results of Experiment 2 stand. However, the results have

⁴¹ The average difference in rating score was 0.83 in Experiment 2 for the first 12 subjects, and 0.5 in Experiment 3; a difference of 0.33.

limited generalizability since the targets were generated from within the face space model. The next experiment (Experiment 4) is designed to explore this shortcoming.

Experiment 4: Celebrity Targets

Experiment 2, and the follow-up study (Experiment 3), served to demonstrate promise for an evolutionary photofit system using a holistic face model. The potential of the system to generalize beyond the face model is tested in this experiment by exploring performance with the target obtained externally. To maximize comparability between studies, the origin of the target is the only change planned.

Method

Experiment 2 and Experiment 3 used targets that had been generated from within the face model. These targets were themselves new or *novel* as they were not part of the original set used to construct the model. It was discussed previously that this approach serves to explore the ability to locate a face that is known to exist within the face space (as the face had been generated from within the model in the first place) but is not informative about the *generalizability* of the model. Namely, is the model capable of generating *any* Caucasian male face? Of course, in a forensic setting, photofits are constructed from facial features not pre-defined in the photofit system. Testing with an *unknown* or *novel* face in this way is therefore more realistic. This experiment evaluates the ability of the photofit system under construction to achieve this goal. Once again, the Face Evolver would be used in the form of an exhibit (in the Hatton Gallery, Newcastle University) and the design should be appropriate for members of the public. Data could once again be collected.

A question arises as to how familiar the target faces should be to someone creating a photofit: should the target face be familiar or unfamiliar? Of course, a familiar face is someone who is either personally known (e.g. a friend) or a famous person (e.g. a celebrity). It is generally believed that familiar faces enjoy a special status in human face perception (e.g. Bruce, 1988; Burton, Wilson, Cowan & Bruce, 1999; Bruce, & Young, 1986; and Ellis, Shepherd & Davies, 1979). In particular, Ellis, Shepherd & Davies (1979) showed that the memory of the internal features of a familiar face is significantly greater than for a face of limited exposure. In addition, for briefly encountered faces, it is the external features – especially the hair – that tend to be preferentially remembered. It was decided that this issue could be resolved at this stage by continual presentation of the target (as before). Differences in familiarity of the targets should therefore be eliminated if the target is always present for reference. System performance with a target evolved from memory could be the focus of a future study.

It was decided to use generally well-known or famous faces as targets. This has the advantage that the evolved photofits could be given to different subjects for identification and

provide a method of system evaluation in addition to rating scores. It was also anticipated that the use of famous faces was likely to be more interesting to participants than evolving someone unknown.

Six famous faces in a full-face pose were located (on the Internet). These were scaled to 180x240 pixels and converted to monochrome. These are shown below -

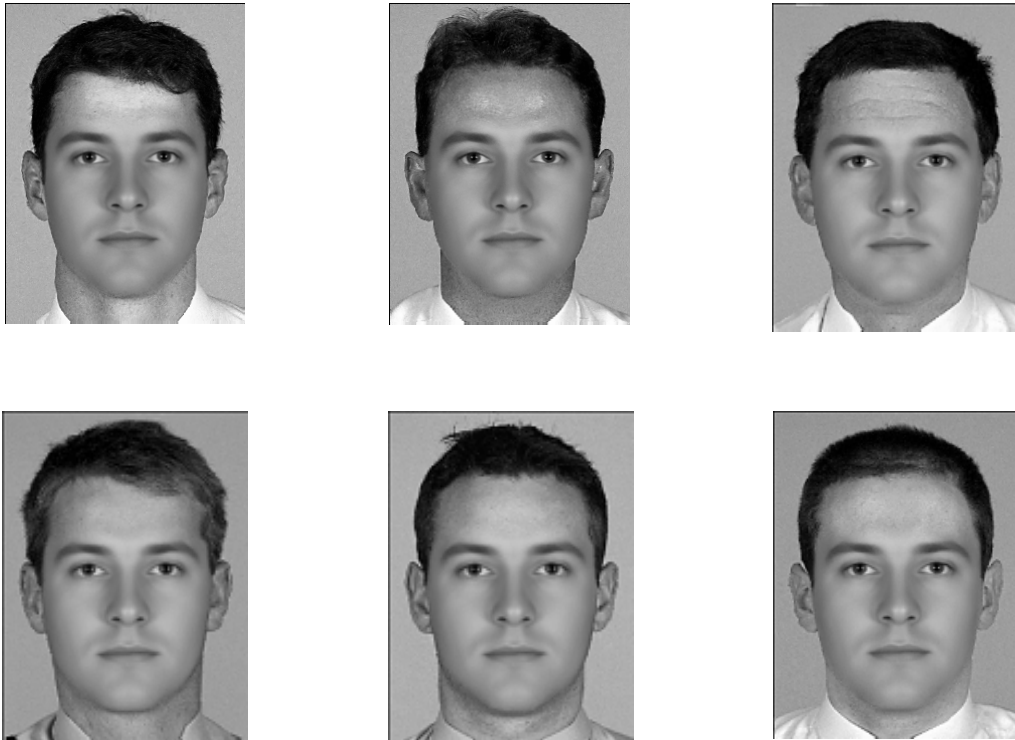
Figure 31: Famous Male Celebrities used as Targets. These are (left to right, top to bottom): Robbie Williams, Tim Henman, Pierce Brosnan, Robert Carlyle, George Clooney and Hugh Grant.

Recall that in Experiment 2, the target's hairstyle was automatically chosen for each of the population faces. This was not possible for the celebrity targets, as the *exact* hairstyle does not appear in the training data. It appeared easiest to make available the 35 hairstyles from the training set for selection; a so-called hairstyle palette. Now, in the two major photofit systems, EFIT and PROfit, each facial feature is selected by swapping out the current selection from the composite under construction. That is, features are always seen and modified in the context of a whole face. This is considered appropriate as it is beneficial to the recognition of individual features (e.g. Bruce, Healey, Burton, Doyle, Coombes & Linney, 1991; Davies & Christie, 1982; Homa, Haver & Schwartz, 1976; Tanaka & Farah, 1993; and Tanaka & Sengco, 1997). A similar approach was adopted therefore.

Although one could make available the original database images for selection of the hair, this was not permissible as identity would be inappropriately revealed. A simple alternative was to key the external facial features onto another face – like the method adopted in Experiment 2 to add the external features to the internal features. The best approach appeared to be to use the internal features from the database's average shape and textured face. This face would be correctly unrecognizable as any of the originals but providing an appropriate context in which to select the hair.

As discussed in Experiment 2, the physical dimensions of the monitor permitted 18 faces of this image resolution to be displayed together. Therefore, 2 screens would be required to display all the available hairstyles. An example of the first six hairstyles keyed onto the average shape and textured face used can be seen in Figure 32 -

Figure 32: The First Six Hairstyles Available for Selection



Participants

Nineteen members of the public visiting the Hatton Gallery, Newcastle University participated. There were 11 males and 8 females and their ages ranged from 17 to 37 (though 7 participants did not complete their age; it was an optional demographic request). 7 participants continued evolution beyond the fourth generation.

Apparatus

The apparatus was the same as in Experiment 2, except that the 6 target faces were obtained on the Internet.

Procedure

The experimental procedure was the same as Experiment 2, except that initially a target was selected from a list of six celebrities (Figure 31), followed by a hairstyle (Figure 32). Participants were able to change the hairstyle at any time during the experiment (a button was present to permit re-selection); the current set of population faces were re-computed to reflect any changes in hairstyle. Evolution continued as in Experiment 2 with the best face being selected first, followed by 5 others from a population of 16 faces. Either 4 or 8 generations were

evolved per person and rating scores of the best faces were recorded. The (famous) target face was on display in the centre of the screen as before during the selection of population faces and whilst rating against the best face.

Results

It can be seen in Table 3 that Robbie Williams and George Clooney were the most selected celebrities for evolution -

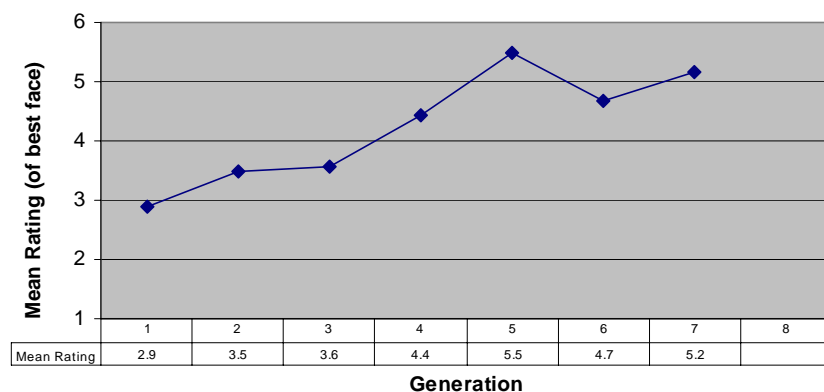
Table 3: Distribution of Celebrities Chosen for Evolution

Robbie Williams	Tim Henman	Pierce Brosnan	Robert Carlyle	George Clooney	Hugh Grant
5	2	3	1	5	3

Examination of the rating data revealed an inconsistency for participant 5. The ratings were 10, 9, 8, 2, 10, 10, and 2. The participant was clearly experiencing difficulty with the rating procedure (or not being serious about the task) since a rating from 10 to 2 on successive generations is inconceivable since the best face is always presented in the next generation. It was considered best that data from this person be removed from further analysis. The remainder of this section considers only 18 subjects, 6 of whom continued to the last available generation. Note also that an unfortunate collection difficulty resulted in the loss of participant ratings for the eighth generation; analysis can only be considered for the first 7 generations in this experiment therefore.

Figure 33 below shows the rating scores averaged for all celebrities and participants -

Figure 33: Mean Rating Scores for the 6 Celebrities (data from Subject 5 omitted)

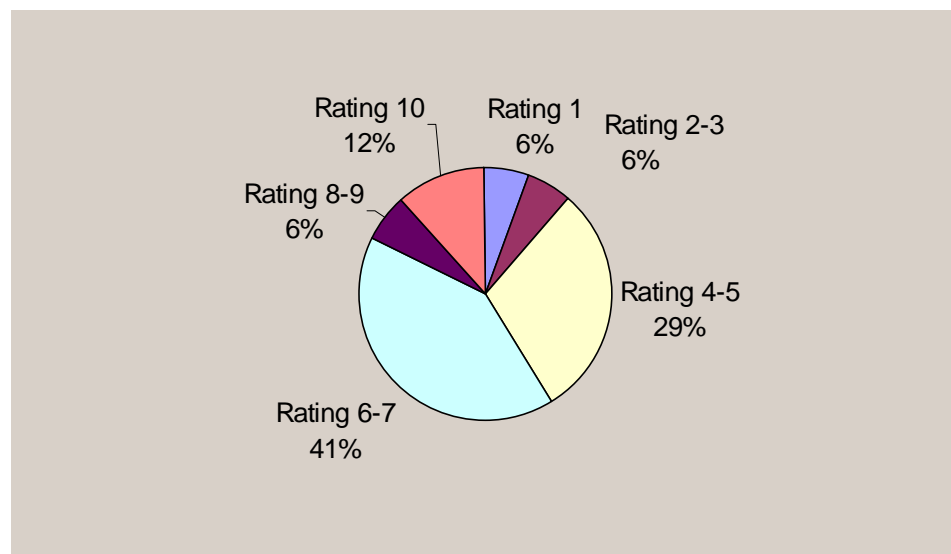


Looking at the graph, it can be seen once again that the average rating scores generally increased with increasing generation. A repeated-measures ANOVA for rating scores for generation 1 to 4 was found to be significant ($F=5.29$, $DF=(3,51)$, $p=0.034$). A t-test, revealed

that rating scores in generation 4 (mean 4.32) were significantly higher than generation 1 (mean 3.26; $t=2.44$, $DF=17$, $p=0.026$); for simplicity, no other analysis were performed over this range. Although analysis is possible for rating scores above four generations, the low number of participants (6) suggests this unwise.

Consider the distribution of the highest score attributed to the best face for each participant. It can be seen from Figure 34 that during evolution, one of the best faces most often fell into the category of “Many Similarities” (6-7 rating bracket). Also about 30% of the time, a face fell into “Some Similarities” (4-5 rating bracket). Interestingly, one can see that about 60% of the time, the peak rating was in the category of “Many Similarities” or higher ($41\% + 6\% + 12\% = 59\%$) and also 12% of the time an “identical” face match was reported. The average peak rating scores were less than in Experiment 2 though. Recall that 95% of the time one face was rated as “Many Similarities”, much more than the 60% reported here. Considering results from the first 4 generations⁴², this experiment (mean 3.6) was found to have significantly lower average ratings than Experiment 2 (mean 4.9; $t=6.15$, $DF=158$, $p<0.05$).

Figure 34: Distribution of the Highest (Peak) Rating for Each Participant

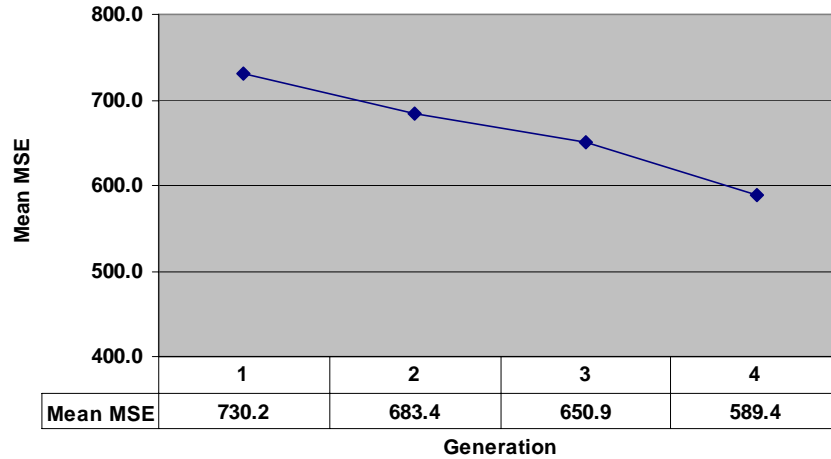


The average MSE score was computed for each generation (with respect to the relevant target) for subjects that evolved for the first 4 generations and is shown in Figure 35. As in Experiment 1, the average MSE scores reveal a general trend of decreasing error with increasing generation. A repeated-measures ANOVA for the MSE data in generations 1 to 4 was found to be significant ($F=14.40$, $DF=(3,51)$, $p<0.001$). Using a t-test, there was a significant

⁴² Analysis would have been skewed should data have been included from generations 5 to 8. This is because there are different numbers of subjects continuing to the eighth generation in this experiment (6 subjects) and Experiment 2 (10 subjects).

decrease between generation 1 (mean 730.2) and 4 (mean 589.44; $t=5.89$, $DF=17$, $p<0.001$); no other tests were computed.

Figure 35: Decrease in MSE Scores with Increasing Generation



Discussion

It was expected to find the trend of increasing rating scores (Figure 33) with increasing generation (as in the previous experiment) and it was satisfying to see that this increase was statistically reliable. With this limited system, it was also pleasing to see that most of the time (59%), participants believed that one of their population faces exhibited at least “Many similarities” to the target face.

Several factors are likely to account for the significantly lower average and peak rating scores of the celebrities compared with the randomly generated targets. The most obvious is that the targets for Experiment 2 were generated from the face model, so that it is clearly possible to produce an exact match. As discussed previously, only 35 faces were used to construct the face model and this limited number may not be sufficient to create an acceptable likeness in general. Note that other PCA studies have used more faces in their corpora (Blanz & Vetter, 1999; Brunelli & Mich, 1995; Kirby & Sirovich, 1990; Sirovich & Kirby, 1987; and Troje & Vetter, 1996), suggesting that a larger database might be preferable.

Irrespective of model size, the fact remains that the number of hairstyles available for selection was rather limited (35). Clearly, hairstyle is of significance, with lower ratings likely to be attributed when differences can be seen in this feature. Note, in this simple task, participants were not instructed to ignore hairstyle changes while rating. In general, there appears to be greater variation, in not just hair, but head pose, head size, lighting and expression than in Experiment 2. Such variations are likely to reduce the subjective quality of the match, resulting in lower recognition rates (e.g. Bruce, 1982; Bruce, Healey, Burton, Doyle,

Coombes & Linney, 1991; Bruck, Cavanagh & Ceci, 1991; Davies & Milne, 1982; Hill & Bruce, 1996; Krouse, 1981; and Wagenaar & Schrier, 1996). Despite these observations, the significant increase in rating scores over 4 generations is an encouraging result for the Face Evolver. The next experiment serves to explore how well these evolved celebrities are recognized by others.

Experiment 5: Recognition of Evolved Celebrities

Analysis of Experiment 4 reveals a significant improvement in the quality of the evolved faces. The acid test of system performance however is the ability to recognize the evolved famous faces. This study therefore tests the ability to recognize the most highly rated faces.

Method

The objective of this experiment was to establish how well people's attempts at evolving the celebrities [from Experiment 4] could be recognized by other people. For the same reasons as Experiment 2, it was decided to select the highest rated face from each subject as the "photofit"; if two faces were found to have the same peak rating, the one in the higher generation was preferred.

As a small pilot study revealed considerable difficulty in recognizing the photofits, it was thought to be of little value to test just the spontaneous or un-cued recognition performance. There are a number of methods that can make such a task easier for subjects. These include the presentation of a semantic cue (e.g. "this person is an American Actor"), a multiple-choice of possible identities, and tasks that match the photofit with the original target. The simplest approach appeared to be to present a list of possible famous people for selection should naming prove fruitless. This has the advantage that the same list could be used for each of the six famous faces. As a further simplification, a "target present" condition was adopted such that the famous person always present for selection in the list. There is considerable justification forensically for the so-called "target absent" arrays (Malpass & Devine, 1981) but this will not be addressed here. It was also decided that a number of other famous people should appear as "foils" so as not to make the task too easy. Ultimately, ten foils were found, comprising an arbitrary selection of other white, male celebrities known in the UK. The following table (Table 4) is a complete list of target names plus foils -

Liam Neeson
Tom Cruise
Robert Carlyle (*)
Ewan McGregor
Robbie Williams (*)
Ronan Keating
Pierce Brosnan (*)
Brad Pitt
Hugh Grant (*)
Michael Owen
Brian Adams
Timothy Dalton
Tim Henman (*)
John Travolta
George Clooney (*)
David Duchovny

Table 4: List of Celebrities Used in the Forced-Choice Task (an asterisk * indicates celebrities that were targets in Experiment 4, the remaining items are foils)

Recall that all except George Clooney were selected for evolution more than once. Now, if all photofits from all Experiment 4's participants were shown to subjects, it is likely that "cueing" would occur (even if the faces were shuffled) and recognition performance might be artificially inflated. As much as possible, it was thought best to avoid this effect. The easiest way appeared to create several "packs" of testing materials, each containing a photofit from one celebrity. Subjects could be (randomly) allocated to one of these test packs. In the end, 4 test packs were created, containing 4 randomly chosen highest rated evolved faces for each famous person.

Participants

Thirty-two visitors to the MacRobert Centre at University of Stirling participated. No demographic information is presented (refer to Experiment 3). Each was approached and asked if they would like to participate in a study involving the naming of famous faces. Participation was voluntary.

Procedure

Each participant was randomly allocated to one of the four test packs. The six photofits in the pack was shuffled before use. The first photofit was presented to a participant and they were asked to identify it (the un-cued naming task). The celebrity list was then presented and the participant asked to select who they thought was the likely target (the multiple-choice task). If the person was still incorrect, they were informed as to the correct identity⁴³ at this point.

Results

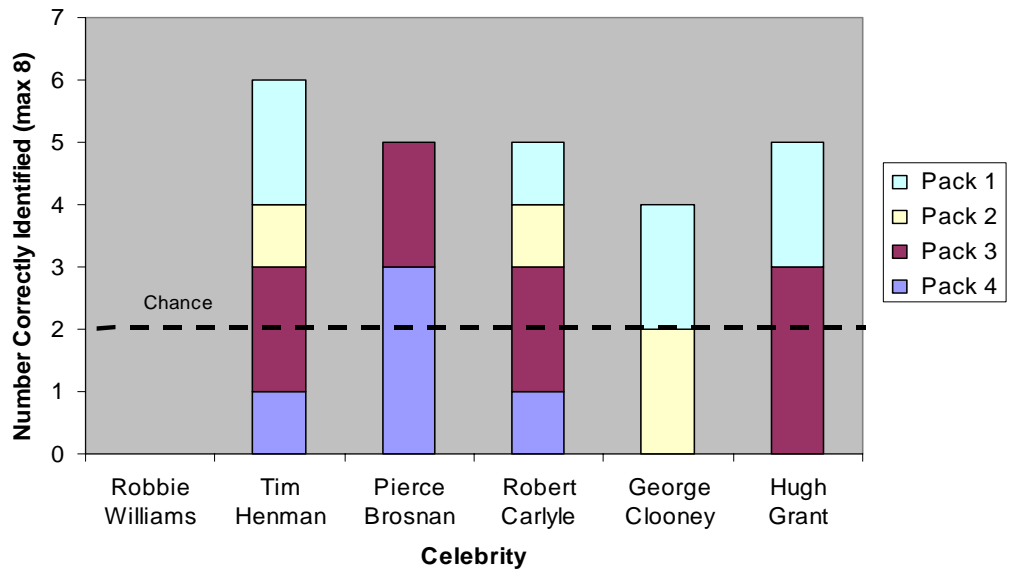
No one guessed the correct identity of any photofit in the un-cued naming task. For the multiple-choice task, 25 celebrities were correctly identified out of 192 possible attempts (13.02%⁴⁴); chance performance was 6.3%⁴⁵. The distribution of correct identifications is shown in Figure 36. The figure reveals that Robbie Williams was not picked out of the list at all, while the other five were all at least double the chance level of 2 identifications⁴⁶; this pattern of observations was significantly above chance using a Chi-Square test ($X^2=42.5$, $DF=5$, $X^2_{crit}=11.07$). The figure also reveals that there were 5 occasions when a face was recognized once and 9 occasions when a face was recognized two or more times.

⁴³ At this point in the design, it seemed interesting to see how people might rate the photofits from memory without seeing the original targets and then with original as a guide. These two tasks were carried out in this order after the multiple-choice task. The data was collected but was put aside for future analysis. ⁴⁴ As there were 32 participants, each shown 6 faces, this resulted in 192 possible identifications in the forced-choice task. $25/192 * 100\% = 13.02\%$.

⁴⁵ As there were 16 celebrities in the forced-choice list, chance level was 1/16 or 6.25%.

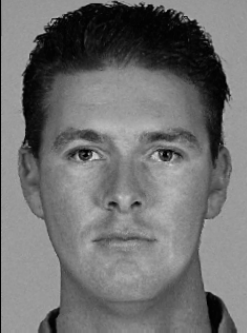


⁴⁶ Chance level was 6.25% x 32 subjects = 2.0.

Figure 36: Distribution of Correct Guesses in the Forced-Choice Identification Task



Examples of the most frequently identified photofits (Figure 37) and those not identified at all (Figure 38) can be seen below -

Figure 37: Examples of the Most Frequently Recognized Evolved Targets (identification at least 2/8 or 25%; ratings assigned during evolution)

Celebrity	Target	Evolved Target
Pierce Brosnan		 <p data-bbox="1038 734 1197 815">Rating = 7 Identified 3x</p>
Hugh Grant		 <p data-bbox="1038 1167 1197 1247">Rating = 7 Identified 3x</p>
Tim Henman		 <p data-bbox="1038 1603 1197 1684">Rating = 6 Identified 2x</p>



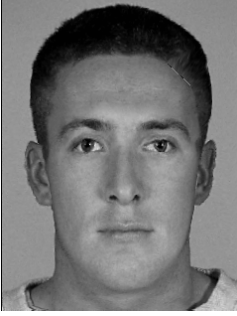

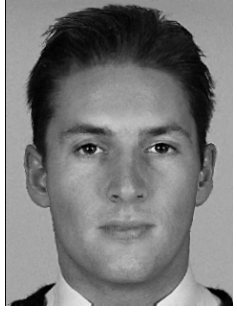


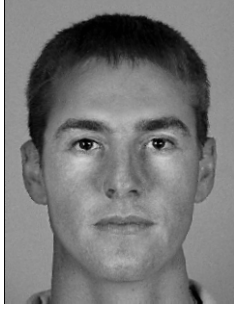


Robert Carlyle		 <p data-bbox="1037 526 1197 604">Rating = 1 Identified 2x</p>
George Clooney		 <p data-bbox="1037 963 1197 1041">Rating = 4 Identified 2x</p>

Figure 38: Unsuccessfully Identified Evolved Targets

Robbie Williams	 Rating 7	 Rating 8
Pierce Brosnan	 Rating 5	 Rating 6
George Clooney	 Rating 6	 Rating 6
Hugh Grant	 Rating 6	 Rating 7

It can also be seen that the most frequently identified celebrity (those that were selected 3 times) was Pierce Brosnan (Pack 4) and Hugh Grant (Pack 3). A low correlation was not significant ($r=-0.28$; $F=1.20$, $DF=23$, $p=0.296$) between the rating scores (in Experiment 4) and the identification rate in Experiment 5.

Discussion

Although none of the targets were spontaneously identified, the overall identification rate of 13% was encouraging from the multiple-choice exercise, especially considering the lack of supervision that participants received during the evolution process. This figure is comparable with the identification rate of 12.5% found from Ellis et al.'s (1975) work, though in that study Photofits were created from memory and selection was made from a target array of 36 composites rather than a list of 16 names. It was also interesting to observe that about two-thirds of the time (64%) when a face was identified, it was identified by at least one other observer. This indicates that when a face is recognized, it is recognized well (25% of the time or more).

It was observed that Robert Carlyle received poor ratings in Experiment 4 (i.e. all ratings were the minimum possible (1) from a single participant). However, the evolved face tested in Experiment 5 was identified (5/32) 15.6% and indicates that the participant was experiencing problems with the use of rating scales rather than in the face evolution process. Once again, this stresses caution with the ubiquitous use of rating scales for this research.

A major problem with the photofits is believed to concern the hairstyle used during evolution. Recall that none of the hairstyles used was an exact match – as was the case for Experiment 2. The possible effects of hairstyle on identification will be explored in the following experiment.

Experiment 6: Appropriacy of the Hairstyle

Recall that in Experiment 5, the photofits were not recognized spontaneously and the identification rate was only 13% when matching faces to a list of names. Part of the reason for these deficits is believed to lie in the chosen hairstyle. If it is the case that changes in hairstyle result in a decrement in recognition - as suggested by Cutler, Penrod & Martens (1987), Hill & Bruce (1996) and Walker-Smith (1978) - then better quality hairstyles may result in better matching. This notion can be explored by obtaining rating scores for the celebrity photofits and testing for a significant difference in ratings between those photofits that were matched and those that were not. It is expected that rating scores would be significantly higher for the photofits that were successfully matched in Experiment 5.

Method

The AFSS used for evolution (Table 2, this chapter) was proposed as a basis for hairstyle rating since this scale has been used considerably and was considered reliable. A new scale (Table 5) was created by changing the word “face” to “hairstyle” in the AFSS.

Table 5: The Anchored Hairstyle Similarity Scale

1	Very poor likeness between hairstyles
2 or 3	Few similarities
4 or 5	Some similarities
6 or 7	Many similarities
8 or 9	hairstyles could be easily confused
10	hairstyles are identical

Simply, the photofits could be shown to a group of independent subjects and asked whether the hairstyle is believed to be “appropriate” for the relevant celebrity. Working from memory in this way has the advantage of avoiding “picture matching” (as might be the case if the original celebrity target was shown along with the celebrity) whilst allowing stereotypical hairstyles for that celebrity to be expressed. Two groups were employed: photofits that were not matched (Group A) and photofits that were matched twice or more (Group B) in Experiment 5.

Participants

Ten participants agreed to complete the experiment, comprising of members of the Psychology Department, University of Stirling. Participation was voluntary.

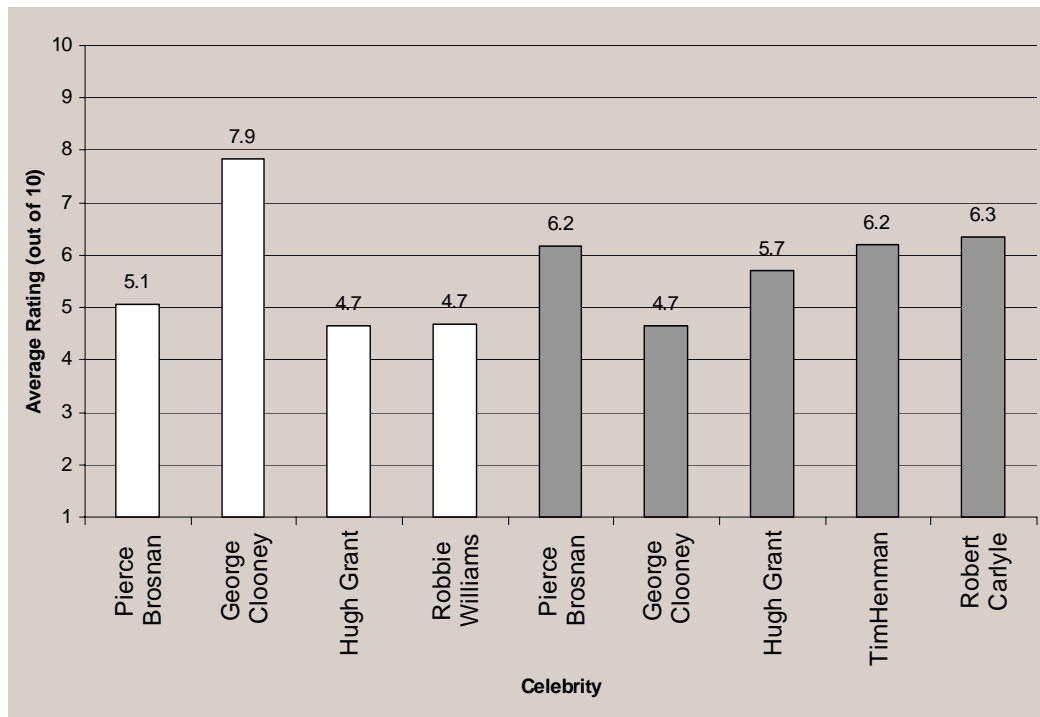
Procedure

The rating scale (Table 5) was shown and described to the participant. Each photofit was then presented to the participant for rating. The twenty photofits were shuffled before each trial.

Results

The overall mean rating was 5.56 (SD 2.15). The mean rating for hairstyles was 5.52 (SD 2.10) for Group A and 5.61 (SD 1.80) for Group B; this difference was not significant using a t-test ($t=0.976$, $DF=184$, $p=0.516$). The reason for the lack of significance between conditions becomes apparent when one considers a plot of the average ratings grouped by celebrity. From Figure 39, it can be seen that the average ratings for George Clooney are outliers: appearing much higher than the other celebrities in Group A and (less so but) much lower than the mean in Group B. Evidence that the George Clooney photofits were outliers is indicated by (1) the rating scores for George Clooney being significantly greater for Group A (mean 7.85) than Group B (mean 4.65) using a t-test ($t=6.40$, $DF=20$, $p<0.001$); and (2) when the data from George Clooney is excluded from the analysis, rating scores for Group B (mean 5.98) are significantly greater than for Group A (mean 4.76) using a t-test ($t=3.73$, $DF=144$, $p<0.001$).

Figure 39: Hairstyle Ratings for each Evolved Celebrity (Group A in white bars, Group B in grey bars)



Discussion

It appeared initially that there was no significant effect of hairstyle rating in this experiment with the ability to match a photofit (to a name) in the previous experiment (Experiment 5). A closer inspection revealed outlying rating data from the George Clooney photofits. His photofits exhibited a “swamping” effect on the rating scores with respect to the other data: the effect size of the George Clooney photofits were large (3.2), significant and in the opposite direction. For this reason the rating data from George Clooney’s photofits were treated separately. Then, it was found that the average rating scores were significantly higher for photofits that were relatively well matched compared with photofits that were not matched at all. This demonstrates that hairstyle is an important feature.

It was also demonstrated is that sometimes a poorly recognized photofit can have what is believed to be a very good hairstyle. The other interesting finding was that the overall rating for hair was 5.56. This corresponds to a rating between “Some similarities” and “Many similarities”. This is generally rather poor and is believed to contribute to the floor level spontaneous naming and the above chance but still poor identification from a list of celebrity names (13%).

The experiments in this chapter have tested face evolution with population of 16 faces. The following experiments are designed to test the effect of increasing the number of faces furthermore: Experiment 7 employs randomly generated targets while Experiment 8 employs

celebrity targets. It is expected that performance should be better for a larger population (given that there is more chance of evolving a fitter individual with a larger population size). For simplicity, the analyses of Experiment 7 and Experiment 8 will consider just the average rating scores of the best faces and compare these data against those obtained from Experiment 2 and Experiment 4.

Experiment 7: Increasing the Population Size, Random Targets

This experiment explores the performance of the evolutionary system with a larger population size. More specifically, it will be designed to replicate Experiment 2 with just the number of faces being increased. This will enable a direct investigation of the effect of population for face evolution. With more faces available for selection, it was expected that the population variability should be greater and evolutionary success higher. In other words, rating scores should be significantly higher than the previous studies that employed smaller populations.

Method

Ergonomically, it was thought best to simply double the number of population faces from 16 to 32; a significantly large increase, it was thought, but not too large to overwhelm a participant. 32 was considered an upper limit at this point since any more computational time might be too irritating for a participant; a population of this size took about 10-11 seconds to generate (using the same computing resources). Simply, this would mean adding another screen to contain the second set of 16 faces and a pair of buttons to navigate between screens. Lastly, it was decided that a simple check be introduced that ensured that the second screen was visited before breeding commenced.

Participants

Eighteen members of the public participated in the study, set up as an exhibit during the Science of the Face exhibition at the Hatton Gallery, Newcastle University (June-July 1999). There were 8 males and 10 females. Their ages ranged from 14 to 65. There were only 4 who chose to continue evolution to generation 8. Participation was voluntary.

Results

The data followed the general trend of increasing rating scores with increasing generation (refer to the combined data chart, Figure 41). A repeated-measures ANOVA for rating scores for generation 1 to 4 was found to be significant ($F=8.41$, $DF=(3,30)$, $p=0.007$).

Using a within-subjects t-test, there was a significant increase between generation 1 (mean 4.32) and 4 (mean 5.94; $t=2.44$, $DF=17$, $p=0.026$); for simplicity, no other analysis were performed over this range. Although analysis is possible for rating scores above 4 generations, the infrequency of participants (4) once again suggests this unwise.

Experiment 8: Increasing the Population Size, Celebrity Targets

Method

The experiment was set up the same as Experiment 4, except that the population size was increased to 32 faces. As another 6 famous faces was available, the number of celebrities was increased to 12: Brad Pitt, Bryan Adams, John Travolta, Timothy Dalton, Alec Baldwin and Ewan McGregor were added. This was believed to help increase the level of interest in the exhibit. The additional target images used can be seen in Figure 40 below –

Figure 40: Additional Celebrities used For Experiment 8. These are (left to right, top to bottom): Brad Pitt, Bryan Adams, John Travolta, Timothy Dalton, Alec Baldwin and Ewan McGregor

Participants

Thirteen members of the public participated in the study, set up as an exhibit during the Science of the Face exhibition at the Hatton Gallery, Newcastle University (June-July 1999). There were 5 males and 8 females. Their ages ranged from 20 to 79. There were only 5 who chose to continue evolution to generation 8. Participation was voluntary.

Results

With the exception of Pierce Brosnan and Alec Baldwin, all celebrities were selected as targets for evolution at least once. The data follows the general trend of increasing rating scores with increasing generation (Figure 41). A repeated-measures ANOVA for rating scores for generation 1 to 4 was found to approach significance ($F=4.28$, $DF=(3,11)$, $p=0.063$). Using a within-subjects t-test, there was a significant increase in rating scores between generation 1 (mean 3.42) and 4 (mean 4.58; $t=2.55$, $DF=11$, $p=0.027$); for simplicity, no other analysis were

performed over this range. Although analysis is possible for rating scores above 4 generations, the infrequency of participants (5) once again suggests this unwise.

Comparison of Chapter Experiments

The effect of generation on rating for Experiment 2 (Random Target, 16 Population Faces), Experiment 4 (Celebrity Target, 16 Population Faces), Experiment 7 (Random Target, 32 Population Faces) and Experiment 8 (Celebrity Target, 32 Population Faces) is shown in Figure 41. The figure clearly shows that there is a general trend of increasing rating with increasing generation. It can be seen here, and also in Figure 42 for overall means, that rating scores were higher for random targets compared with celebrity targets, and also where the population size was increased from 16 to 32; data compiled from the first 4 generations only. The random targets (mean 5.77) were rated significantly higher than the celebrities (mean 4.18; $t=6.69$, $DF=372$, $p<0.001$); a difference of 1.59. The faces evolved from the population of 32 faces (mean 5.52) were rated significant higher overall than for a population size of 16 (mean 4.75; $t=3.16$, $DF=372$, $p=0.002$); a difference of 0.77. The random targets with 32 population faces (mean 6.27) were rated significantly higher than the random target population of 16 (mean 5.35; $t=3.13$, $DF=214$, $p<0.001$); a difference of 0.92. The celebrity targets with 32 population faces (mean 4.54) were rated significantly higher than the celebrity target population of 16 (mean 3.90; $t=3.13$, $DF=214$, $p<0.001$); a difference of 0.64.

Figure 41: Comparison of Rating Scores between Experiment 2, Experiment 4, Experiment 7 and Experiment 8

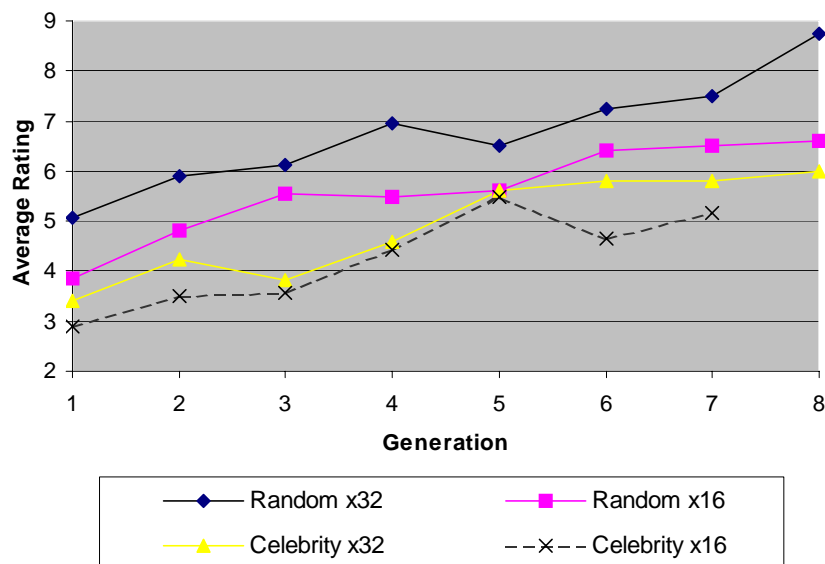
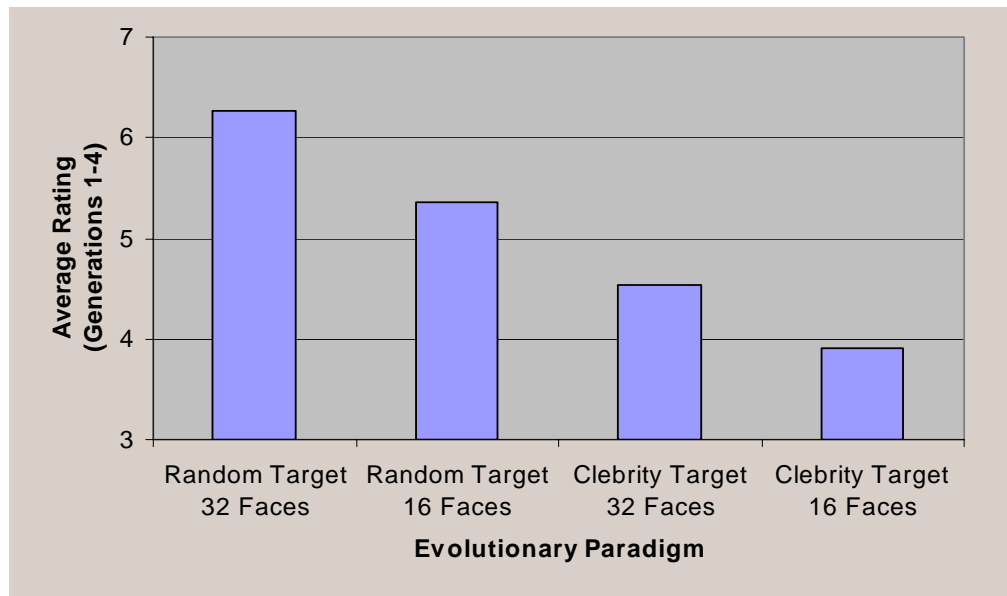


Figure 42: Rating Performance for Target Type and Population Size (data from generations 1 to 4 only)



Discussion

There was a significant increase in rating scores (of 0.77) when the population size was increased from 16 to 32, indicating the importance of larger populations for this type of methodology. An interesting question is whether this could continue further? It appears from De Jong's pioneering work in Genetic Algorithms that the best population size appears to lie in the range of 50-100, though when on-line factors (e.g. the time taken to create a population) are taken into account, there is evidence from Grefenstette (1986) that 30 is an appropriate number of individuals (Mitchell, 1996).

Although the current study employed a population size similar to that of Grefenstette, could the benefit of enhanced performance (over a smaller population) be the consequence of more individuals being present in the *first* generation only, with no added benefit to a larger population for the remaining time? Note that although not mentioned in the results, there is a first generation advantage since the rating scores for generation 1 were significantly higher in the larger population (mean of 4.40) compared against the smaller population (mean of 3.43; $t=2.14$, $DF=68$, $p=0.018$); a difference of 0.98. Anyway, the question posed above can be addressed by comparing rating scores in studies with 16 faces (i.e. Experiment 2 and Experiment 4) and 32 faces in the population (i.e. Experiment 7 and Experiment 8) after subtraction of the rating score assigned in the initial generation.

When this pre-processing is performed, a one-tailed between-subjects t-test reveals an approaching significant increase for the larger population size in generations 2 to 4 ($t=1.51$, $DF=208$, $p=0.066$); an increase of 0.46. This suggests that a constantly larger population size is

of most benefit⁴⁷. Obviously, this finding fits the notion that a larger population throughout the evolution process *should* be beneficial since the diversity in the population would generally be greater, increasing the chances of producing “superior” individuals. Note, however, that an examination of the effect size reveals that most improvement (in an increase in population size) occurred in the *first* generation: the average ratings increased by 0.98 in the first generation but only 0.46 for following four generations; a difference of more than 100% (113%). This suggests that the initial population size is relatively more important than ones that follow.

It is also apparent that, overall, the random targets were evolved significantly better than the celebrities (by 1.59 rating points). As discussed previously, this effect could be due to limitations in model complexity but other factors are known to be important here: differences in lighting, head size, head orientation, facial expression, and poor availability of hairstyles. There is also a potential confounding effect of a participant having to select a hairstyle (and not choosing the most appropriate), the deficiency of methods available for “improving” the quality of the hair anyway (e.g. by modification in a photographic editing facility such as Adobe Photoshop) and a lack of an operator to assist.

General Discussion

The set of four evolutionary experiments (Experiment 2, Experiment 4, Experiment 7 and Experiment 8) found consistent results for the Mark II Face Evolver. Firstly, the rating scores significantly increased from generation 1 to generation 4 for each experiment and the trend of increasing rating was present in the few subjects that continued to the eighth generation. Additional confirmation of this significant increase in rating scores was provided by independent subjects in Experiment 3. A peak rating score, the maximum rating for a evolutionary run, in at least the “Many similarities” category was recorded for 95% of subjects with the random targets with 16 population faces.

These results suggest that the selection of whole faces, rather than via rating scales, is a valid method of user input. In addition, unlike rating scales, there are no known problems with this method of selection (i.e. whole face selection). Of course, there are other schemes that one could adopt. For example, a more fine-tuned approach that ranked the selected faces, as implemented in Baker & Seltzer (1998), might provide a faster convergence. The current version of the Face Evolver did use a coarse ranking scheme such that the most perceptually-similar face (i.e. the best face) received twice the influence of the others. However, the effectiveness of this bias is unknown and is one of the issues explored in the following chapter.

⁴⁷ It is assumed that only an *approaching* significance was found due to an insufficient number of subjects. This is not considered to be of real importance here; it is the trend that is taken to be relevant.

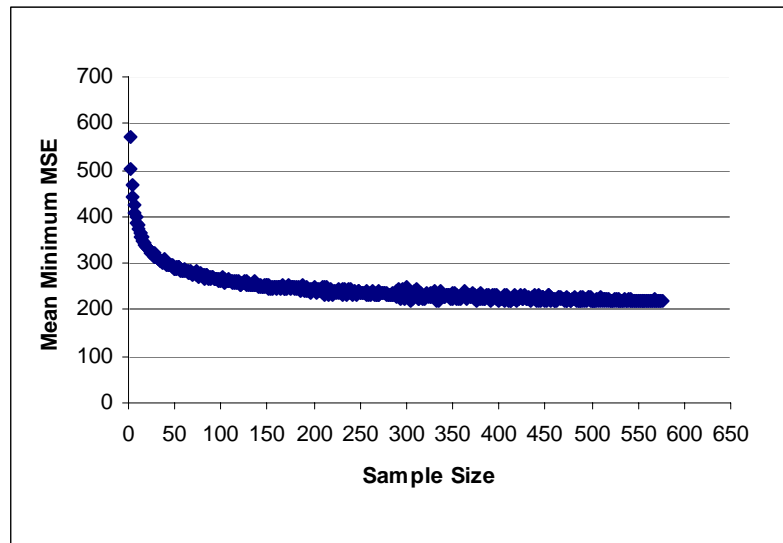
It was not surprising however to see a drop off in performance for the celebrity photofits given that the face model was rather limited. Though, even with 35 faces, the majority of people (60%) reported evolving one of their faces to exhibit at least “Many similarities” with the celebrity. Even so, the “photofits” from Experiment 4 were clearly not sufficiently close to the target for any of the subjects in Experiment 5 to spontaneously recognize them. It was shown that for them to be correctly identified from the list, hairstyle was an important feature. But, in an interesting way: when the data from George Clooney was removed, the relatively well recognized photofits (those with two or more correct in the multiple-choice list) were rated as having a hairstyle significantly more appropriate to the celebrity than the unrecognized ones; in fact the reverse was true for George Clooney. This demonstrates that overall hairstyle is important for identification, though sometimes even a photofit with a relatively low-rated hairstyle can enjoy relatively good recognition. This observation is reflected in a comment left by Participant 2 in Experiment 8, “Matched really well from the first generation ... hair was a problem though.” Note overall, though that subject ratings were quite low (5.55), between the categories of “Some similarities” and “Many similarities”, and may be one factor responsible for spontaneous naming remaining at floor level.

The studies were also able to demonstrate a marked effect of population size. When the number of faces in a population was increased from 16 to 32, the rating scores did not just continue to increase as before, but increased at a greater rate (just falling short of significance). This indicates the value of a constantly larger population, providing considerably more opportunities to produce superior individuals, resulting in faster evolution. When the population size was doubled though, it was curious that a much larger increase in rating scores (over 100%) occurred in the first generation compared with the increase over the other generations. This is presumably because, in this paradigm, an initially closer likeness can be found by the random generation of faces rather than by evolution.

It is possible though to characterize the effect of population size on the best face produced for the initial generation, using the minimum MSE measure. Recall that Experiment 7 involved 32 population faces presented to 18 subjects, providing a pool of 576 different randomly generated faces in the first population. It is possible then to conduct a non-replacement re-sampling of this set with varying population size to explore the quality of the best face. For example, for a population size of 10, a set of 10 population faces would first be selected randomly and the minimum MSE identified. This would be repeated until all complete sets of 10 faces (570 in total) had been sampled (the 6 remainders would be ignored at this stage). The average minimum MSE would be calculated for this population size. The process could then be repeated a number of times (e.g. 10), taking the average, to limit sampling anomalies (e.g. chance inappropriate clustering of relatively good or poor samples) and to take into account ‘remainders’.

The following plot was obtained when applying this analysis to population sizes between 2 and 576 –

Figure 43: Reduction in minimum MSE with increasing population size (randomly



generated faces)

It can be seen that most benefit in an increase in sample or population size occurs over the first 50 faces. This suggests that generating a larger population is unlikely to result in a face that is markedly better than one already generated. A relatively straightforward equation can be computed relating the population or sample size (s) and average minimum MSE (e), explaining 96% of the variance in the data -

$$e = 494 s^{-0.132} \quad \dots (1)$$

The equation also predicts that, compared to a unity sized population, the minimum MSE will have decreased on average by 50% with a population size of 14, 71% for a population size of 50, and 80% for a population size of 100. This suggests that an initial population size of about 50 is a sensible trade off between generating a relatively good best face and not exposing (a subject to) too many facial stimuli.

In addition to the trend of convergence of the best faces to the target, a measure of peak performance, it was found that the average error score of the face population in Experiment 2 became significantly less with increasing generation. This indicates that the population as a whole was becoming more like the target face (i.e. the average pixel intensity of the images approached that of the target). The other observation was that there was a lack of any significant differences in the standard deviation of the MSE. This later finding suggested that there is no evidence that the variation in the population had been “lost” and therefore evolution could continue for more generations, successfully improving the target match.

One influence though that increases variability in a population is the presence of mutation. A small mutation was introduced during breeding to encourage population variability. The parameter was set to a “guesstimated” value of 0.05; though the effect was not known. Indeed, with the exception of the initial population size and gross changes to the

population size, the effect of most evolutionary parameters is not known: is the mutation rate really a useful feature and is 0.05 appropriate? Is *elitism* really effective here? As mentioned above, is it appropriate to use a 2:1 selection pressure for the best face compared with the other selected faces? How many faces should be selected? Is it even more beneficial to increase the size of the initial population still further? These are of course major issues for an evolutionary photofit system.

In addition to establishing optimal parameter settings, there are naturally further developments that need carrying out in order to move closer to a practical photofit system. Two major areas of development have been identified so far. These are both related to the evolution of celebrity faces: better hairstyles and an increase in the complexity of the face model. The hairstyle itself illustrates a necessary change in the method by which photofits need to be created. Up to now, evolution has been done “automatically”, with little external influences, but the hair is a highly idiosyncratic feature and will necessarily involve much detailed work to obtain an acceptable likeness. This is likely to be achieved via a large repertoire of hairstyles followed by manipulation in a photographic editing package. It will be necessary therefore for an operator to be present to “guide” hairstyle creation. Indeed, the use of an operator is likely to be of considerable importance anyway: what should a user do if the set of initial faces are so poor (by chance) to be of little value to a witness? It would be necessary then for more faces to be “created” – a likely job for an operator.

It is proposed though, that before more development is carried out, that effort be spent exploring the appropriacy of the current set of evolutionary parameters. One *could* systematically manipulate each parameter in the experimental paradigm used to date. The normal duration of experiments suggests that this approach is intractable in the time available for the project (i.e. 3 years). An alternative is to run separate simulations to validate these parameters. To achieve these objectives, the following chapter (Chapter 4) explores the parameter space of the evolutionary system and the one after that (Chapter 5) continues system development and considers system performance in more ecological ways (esp. photofitting from memory).

Chapter 4: Simulations

This chapter explores the performance of the Face Evolver through a series of short computer simulations. Each simulation is carefully set up to investigate the effect of each parameter setting. Overall, it is found that the population size, elitism and selection pressure (of the face with the best likeness) have parameters set appropriately in the previous chapter, but that the mutation rate should be increased (to 0.1) and the number of faces selected by a user could be justifiably reduced (to a minimum of three or four). Also, it is found that similar performance results from population sizes in the range of 10 to 32 faces when evolved for an equivalent length of time. In addition, the use of coefficient pruning and the separate selection of shape and texture components appear to make a significant impact on convergence. These findings become valuable for further development carried out in the following chapter (Chapter 5).

Appropriacy of Parameter Settings

The Mark I and Mark II versions of the Face Evolver of the previous chapters have demonstrated encouraging results. These programs were designed with extreme care with the aim of producing optimum results. Parameters were based on what was believed to be "appropriate" settings. But, was 0.05 the best value for the mutation rate? Was six an appropriate number of faces to be selected for each generation? Was a two-to-one (2:1) selection pressure on the best face reasonable?

One way to finalize these parameters would be to run a separate evolution experiment for each. The problem with this approach is that it is exceedingly time consuming. For example, one may wish to try a range of mutation settings to gain a good understanding of this parameter: perhaps from 0.00 to 0.2 in 0.05 steps. This naturally results in 5 separate experiments (including the zero mutation baseline). If each experiment collected data from a sufficient number of subjects to obtain statistically reliable data, for example 20, just this manipulation is likely to take a month to complete. More experiments would be required to understand the role of other parameters.

Such an investigation would tend to restrict further development on the photofit system. What is required is a mechanism to indicate suitable parameter settings on a much shorter time scale. A compromise must of course be made, since it is unlikely that any alternative approach would provide data as valid as that obtained experimentally. The best compromise appears to reside in a set of computer simulations. This approach attempts to provide a model that is as close as possible to one or more aspects of the real world. This model can then be considered "real" and explored experimentally. The advantage is that the model can be computerized and be much faster to investigate than the real world.

To run the Face Evolver system as a simulation, automatic selection of the population faces is required. In GA terminology, each population face is selected via a score or *fitness* obtained by a *fitness function*. The higher the fitness, the greater the influence of a face. Arguably, the simplest fitness function would be derived via MSE scores of faces in the population. It seemed from Chapters 2 and 3 that MSE was a reasonable indicator of system performance (see also Appendix A). Recall that scores became significantly less with increasing generation, indicating that the population faces became on average significantly closer to a target. Used in this way, lower MSE values would indicate greater fitness and a better chance of being selected as a parent. Arguably, a natural consequence of this approach is that the face with the lowest MSE in a population would be selected as the highest similarity face: the so-called “best” face for a generation.

If these notions are valid, then the MSE measure should select the same faces as participants in one of the previous experiments. When this is attempted, it is found that the model correctly predicts 180 out of a possible 348 selections (or 52%) made by subjects in Experiment 2; this is a significant increase from the chance level of 6/16 or 37.5% ($X = 18.78$, $DF=1$, $X_{crit} = 3.84$). The model also predicts 45 out of a possible 348 best selections (13%); once again, this is a significant increase from the chance level of 1/16 or 6.25% ($X = 24.85$, $DF=1$, $X_{crit} = 3.84$). As this simple model predicts performance significantly above chance, it will be used as a basis for simulation. It is understood that it is not a perfect reflection of the pattern of face selection made by subjects. The results are therefore taken as performance “indicators”.

The Simulations

Simulations will begin by looking at the effect of modifying the number of faces in a population: to provide baseline data for the following parameter manipulations. Further simulations will then explore selection pressure, mutation rate, the number of faces selected and elitism. Subsequently, more simulations will be run to re-investigate the effect of population size and two novel approaches: coefficient pruning and a modified selection mechanism. System performance will be assessed by average and peak MSE scores. The programming will be carried out in Matlab, since the main non-windows part of the Face Evolver software is also resident in Matlab (and no complex windows-based interface is required).

Simulation 1: Population Size

The first set of simulations is designed to look at the general effect of increasing the number of individuals used for evolution: the number of faces in the population. These initial simulations are designed with “flat” settings: no mutation, no elitism and unity selection pressure (i.e. the best face has the same opportunity of breeding as the other faces selected).

Following this procedure results in a baseline performance that enables comparison against other parameter manipulations.

It is proposed that the investigation into population size be carried out from 4 to 32 faces in order to gain a good understanding of performance. Further, it is suggested that the number of generations should be increased beyond 8 (used in prior experiments) until a stabilization in performance is reached. For these simulations, this occurs when the MSE scores no longer change. However, in later work, the presence of mutation in the shape or texture coefficients of a face will constantly add variability to the generated faces, resulting in fluctuations in the MSE. Even for small mutation rates, for example 0.01, these fluctuations will be present. Convergence is then assumed to have occurred when the major MSE changes can be seen to have taken place⁴⁸. In practice, it was decided sensible to run the Face Evolver for a large number of generations, many more than would ever be imagined to run in a photofit session with a witness. To this end, 40 generations were run.

It is proposed therefore that system performance be measured via the mean and minimum MSE scores from faces in a generation. These data will provide a measure of average and peak system performance respectively. Of these two measures, it is the minimum MSE that is considered more important since the primary purpose of a photofit system is to create *one* face that can be used as a photofit. Analysis of the average MSE is considered fruitful as lower values indicate a concentration of solutions in locations closer to the target in face space. The result should be an increase in the probability of locating highly fit individuals and be reflected in lower scores in the average minimum data.

Further, to keep the analysis tractable, since data tends to be “semantically rich”, emphasis will be on the *initial*, *early* and *ultimate* (or converged) generations. Note that variations in the initial generation are not relevant to mutation, elitism or selection pressure since these factors affect populations following breeding. Analysis of “early” generations, defined to be the first ten, is especially relevant since this figure is taken as a likely guideline to the number of generations required to achieve a good target likeness⁴⁹.

The simulator would be run for 40 generations, with each run increasing the number of population faces by a small amount: 4, 8, 12, 16, 20 and 32. To gain a measure of consistent performance, each population size was run with a selection of targets. Ultimately, the original

⁴⁸ In the limit, one might expect the evolutionary system to converge on a given target. Due to time constraints, this notion was not investigated since a *trend* of asymptotic performance was observed for small mutation rates in Simulation 3 extending beyond 40 generations. Later work could of course explore performance after a very large number of generations (e.g. several thousand).

⁴⁹ This figure was chosen to match the estimated number of generations (from Chapter 3) for a face to be bred to a rating of 8 or more on the FRSS. This level corresponds to the categories of “Faces could be easily confused” and “Faces are identical.”

35 targets that featured in Experiment 2 were used, with each target being evolved twice. Each trial (i.e. a run of 40 generations for a specified target) would commence with a different set of random faces. As before, this serves to avoid idiosyncratic behaviour occurring when an initial set of faces are either relatively desirable or relatively poor (compared with a target) due to the chance selection of parameter values.

With different numbers of faces in a population, the actual number of faces selected would need to be carefully chosen. If 6 faces were always chosen, as was the case in Chapter 3, then 3/4 of the total faces would be selected for a population size of 8 but only 1/5 if the population size was increased to 32. This means that most of the faces would be chosen for the smaller population but only the most fit would be chosen in larger one. The consequence is that the larger population is likely to converge very much faster because of the preferentially better faces being selected as parents. The chosen solution is to keep constant the proportion of selected to total number of faces. This appropriately results in an equal proportion of fitter individuals being selected for each population size. In keeping with Chapter 3, the fraction 6/16 (or 37.5% of the population size) can be used⁵⁰. This results in 2 faces ($\lceil 1.5 \rceil = 2$) being selected for a population size of 4, 3 faces for a population size of 8, and so on.

Running the simulator over the proposed population range and computing average MSE scores over 70 runs resulted in Figure 44. This plot reveals that there is no difference in the mean MSE scores for the starting generation. This is an expected result because each point is calculating the average of a set of random faces, which will be the same, barring noise.

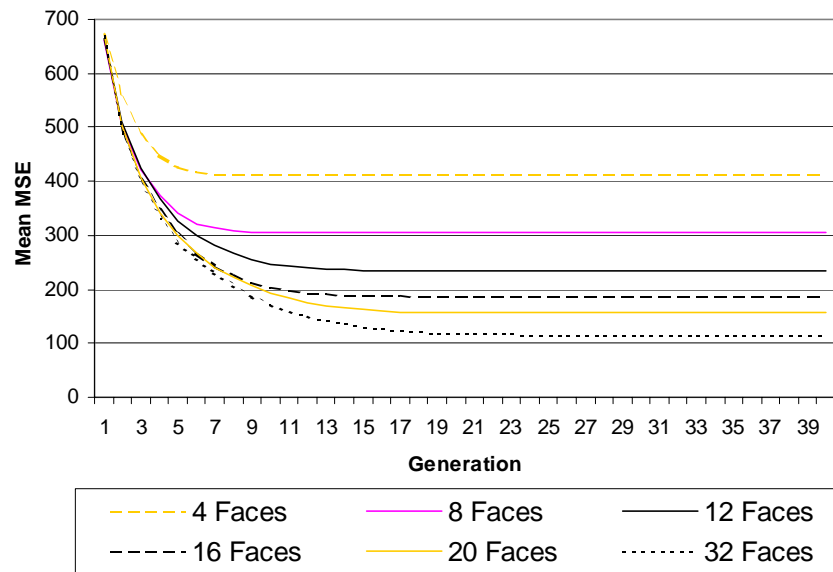
For other generations, there is an initial rapid decrease in the average MSE and this rate of change becomes progressively less over time. Also, it can be seen that increasing the population size increases the time for the MSE to converge (about 6 generations are required for 4 population faces, 14 generations for 12 faces, and 20 generations for 32 faces). Further, note that increasing the population size not only produces faster convergence (for example, a mean MSE of 250 takes about 10 generations for a population size of 12 but only 6 generations with 16 population faces) but also the ultimate MSE is less (from 411 for 4 faces, 235 for 12 faces, and 157 for 32 faces). Over the early generations (from generation 2 to 10), there is no significant difference in the average MSE scores between 12 and 16 faces ($t=0.63$, $DF=628$, $p=0.530$), and 16 and 20 faces ($t=0.35$, $DF=628$, $p=0.726$), but there is a significant difference between 16 and 32 faces ($t=3.31$, $DF=628$, $p<0.001$). Overall, the results suggest that population size is important, but large changes are required (e.g. doubling the population from 16 to 32) if significant increases in performance are to be found.

⁵⁰ Mathematically, the number of selected faces, $N_s = \lceil 6/16 * N_p \rceil$

N_p = number of faces in the population

$\lceil x \rceil$ is the *upper* function of x ; the presence of a non-zero fraction of x will return the next integer, otherwise x is returned. For example, $\lceil 1.001 \rceil = 2$ and $\lceil 6.0 \rceil = 6$.

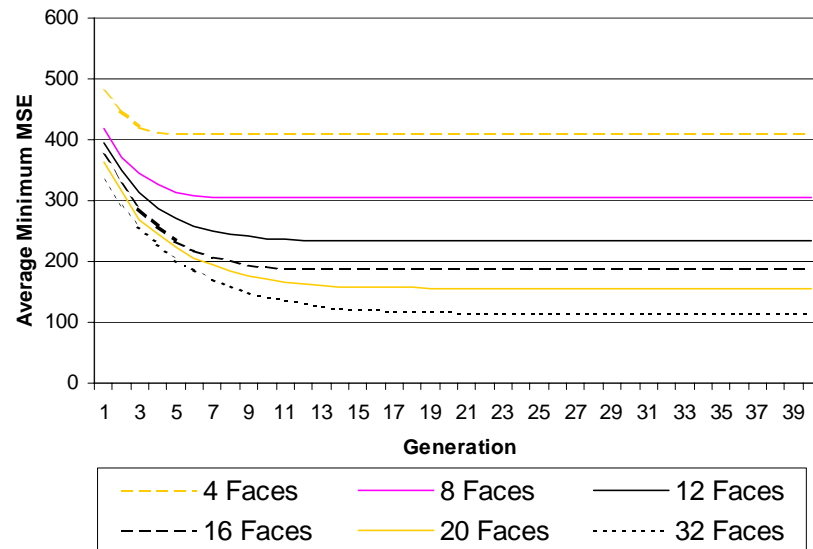
Figure 44: Effect of Varying Population Size on Average MSE (1:1 selection pressure, mutation probability of 0.0, no elitism and $\lceil 6/16 * \text{population size} \rceil$ selected faces)



Turning now to the data for the average minimum MSE. From Figure 45, it can be seen that with increasing population size, the average minimum MSE is less in all generations *including* the first. This graph backs up an observation from Experiment 7 and Experiment 8 that a larger initial population size is preferable since it increases the likelihood of fitter individuals being generated. Much the same as the average MSE data, the rate of change of the average minimum MSE becomes progressively less rapid with time and the terminal value decreases with increasing population size.

These data demonstrate that evolution appears to be working. Consideration of the appropriate number of population faces to use will be made later when suitable parameters for the mutation rate and selection pressure have been established in conjunction with the use of elitism. Selection pressure will be examined first.

Figure 45: Effect of Varying Population Size on Average Minimum MSE (1:1 selection pressure, mutation probability of 0.0, no elitism and $\lceil 6/16 * \text{population size} \rceil$ selected faces)



Simulation 2: Selection Pressure

It is proposed to test the effect of selection pressure of the best face using the parameters of Simulation 1 with 16 population faces. Using this procedure, the six faces with the lowest MSE in a population were attributed a fitness value of 1.0. However, the face with the lowest MSE overall was taken as the best face and assigned a fitness of 2 for one simulation, and 3 for the next: a selection pressure of 2:1 and 3:1 respectively. A further simulation for a unity or 1:1 selection pressure was not necessary since this has already been carried out in Simulation 1 (i.e. the simulation for 16 population faces).

The seed value used to initially set the random number generator was kept the same as in Simulation 1. In addition, this same seed value was used again at the start of simulations with selection pressures of 2:1 and 3:1. This serves to improve comparability of results by ensuring that all simulations begin with the same set of faces.

When the simulator is run, Figure 46 indicates an advantage for progressively higher selection pressures on the average MSE only in the first part of the early generations: up to the first 6 generations (3:1 curve) or 7 generations (2:1 curve). Later generations appear better with unity selection pressure. In contrast, the average minimum MSE (Figure 47) indicates no benefit over generations 1 to 3 and worse performance thereafter for either of the non-unity selection pressures.

For a 2:1 selection pressure, there is no significant difference in the average MSE scores over the early generations compared with the 1:1 baseline condition ($t=1.14$, $DF=628$, $p=0.255$), though there is an approaching significant decrease over generations 2 to 6 ($t=1.88$, $DF=628$, $p=0.061$). In contrast, the average minimum MSE is significantly worse over the early

generations for 2:1 and 3:1 selection pressures ($t=3.31$, $DF=628$, $p<0.001$). At this stage, it appears best if a 2:1 selection pressure is not used in the Face Evolver.

Figure 46: Effect of Varying Selection Pressure (of best face) on Average MSE (16 population faces with 6 faces selected per generation, mutation probability of 0.0 and no elitism)

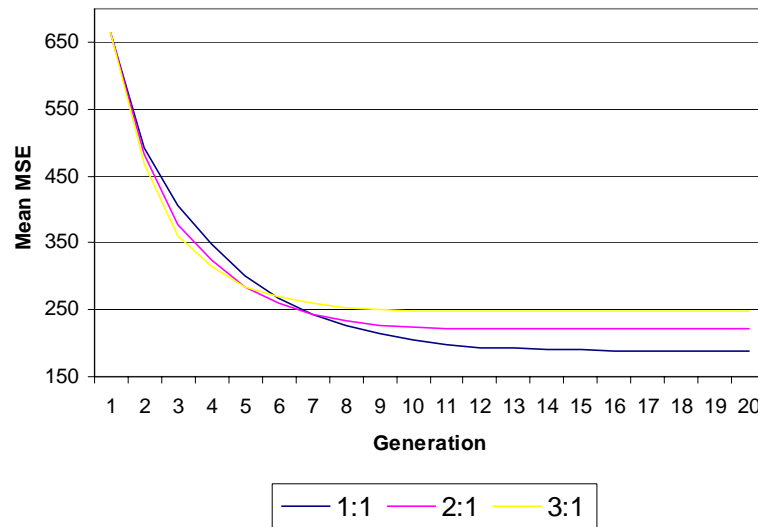
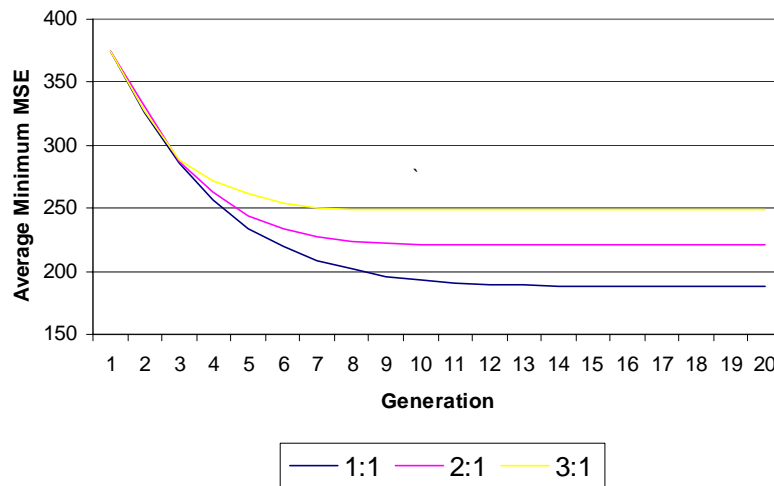


Figure 47: Effect of Varying Selection Pressure (of best face) on Average Minimum MSE (16 population faces with 6 faces selected per generation, mutation probability of 0.0 and no elitism)



Simulation 3: Mutation Rate

It is proposed to test the effect of mutation rate using the parameters of Simulation 1 with a selection pressure of 2:1 (maintained the same as in Experiment 2 at this stage). A range of mutation rates was tried: 0.01, 0.05, 0.1, and 0.2. Figure 48 shows that compared with the

reference (a mutation rate of 0.00), there is no benefit of a mutation rate of 0.2 mutation on the average MSE. The mutation rate of 0.1 only appears valuable after about 20 generations. Of the remaining settings, 0.01 appears to decrease the fastest but the final value appears no better than a 0.05 rate.

On the other hand, the terminal scores (Figure 49) reveal that some mutation is beneficial to minimum MSE performance. The lowest average minimum scores can be seen for rates of 0.05 and 0.1 but mutation settings on either side of this value result in much worst scores. It is the case that too little or too much mutation is undesirable. As the effect of 0.05 and 0.1 rates is very similar, and is non-significantly different over 2 to 10 generations ($t=0.44$, $DF=628$, $p=0.658$), it is proposed at this stage that the value of 0.05 be kept the same as previous experiments in Chapters 2 and 3.

Figure 48: Effect of Varying Mutation on Average MSE (16 population faces with 6 faces selected per generation, no elitism and 2:1 selection pressure)

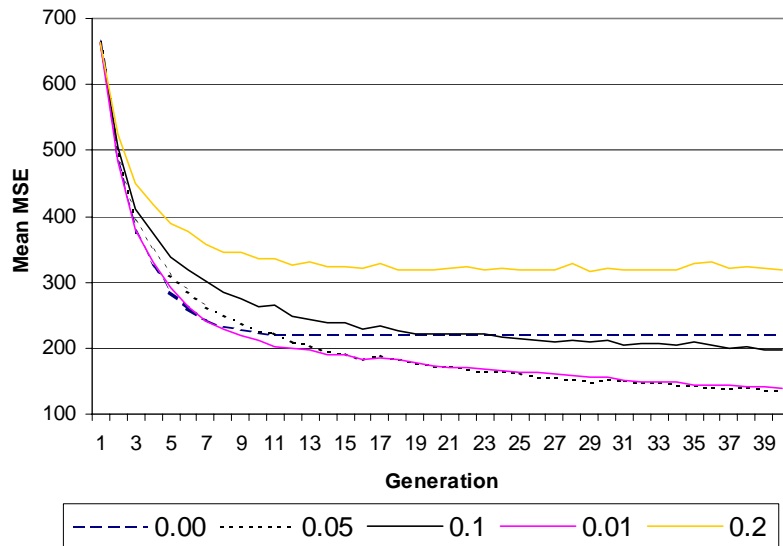
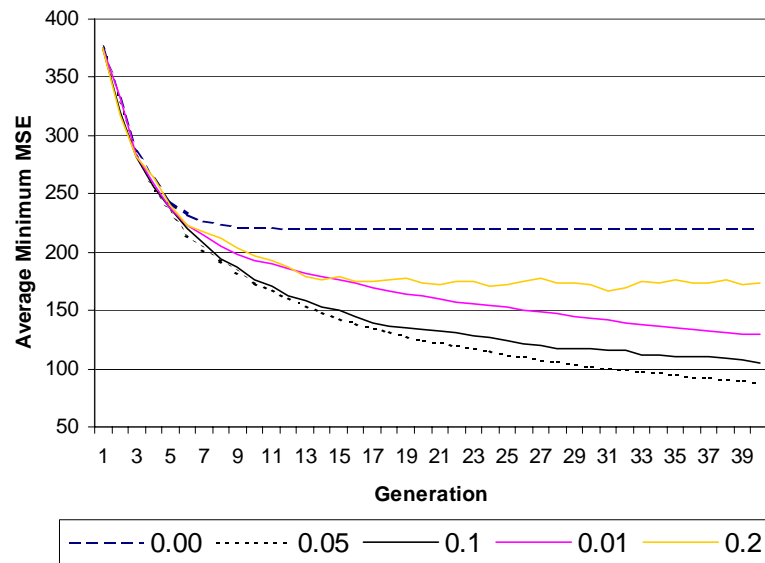


Figure 49: Effect of Varying Mutation on Average Minimum MSE (16 population faces with 6 faces selected per generation, no elitism and 2:1 selection pressure)



Simulation 4: Number of Selected Faces

The next batch of simulations are designed to explore the benefit of selecting different numbers of faces from a population. Recall that six faces were selected in Simulations 2 and 3. The effect will be examined by incrementally increasing the number of faces selected from 3 to 8. A mutation rate of 0.1⁵¹ and a 2:1 selection pressure will be used. Results from the graphs of the average MSE (Figure 50) and average minimum MSE (Figure 51) reveal that reducing the number of selected faces down to 4 is of benefit in the early generations⁵². With 3 faces selected, there is an initial benefit to 6 generations (on both measures) but after this, there appears no difference compared with 4 faces selected for average MSE and worse performance for average minimum MSE. There is a significant decrease in minimum MSE over generations 2 to 10 from 6 to 5 faces selected ($t=1.99$, $DF=628$, $p=0.047$), from 5 faces to 4 faces selected ($t=2.76$, $DF=628$, $p=0.006$), but not from 4 faces to 3 faces selected ($t=0.02$, $DF=628$, $p=0.982$).

Concern was expressed that the selection of as few as 3 population faces might lead to an inappropriate reduction in the population diversity. The result of an ANOVA for the standard deviation of the MSE scores for populations of faces over the 10 generations was not significant ($F=0.56$, $DF=(9,419)$, $p=730$). This indicates that there is no evidence to suppose that the population diversity is significantly different in any of the conditions, including the selection of 3 faces.

⁵¹ This choice between a mutation of 0.05 and 0.1 was arbitrary.

⁵² To simplify the graphical presentation, only the first 10 generations are shown.

The strong suggestion then is that the number of faces should be limited for faster conversion, with 3 or 4 faces being a sensible lower limit. The *exact* number is not considered of great importance – knowledge of the potential inappropriacy of selecting too many faces is the take home message.

For the remaining simulations, the number of selected faces will be kept at six (to maintain comparison between studies).

Figure 50: Effect of Varying the Number of Faces Selected on Average MSE (16 population faces, mutation probability of 0.1, no elitism and 2:1 selection pressure)

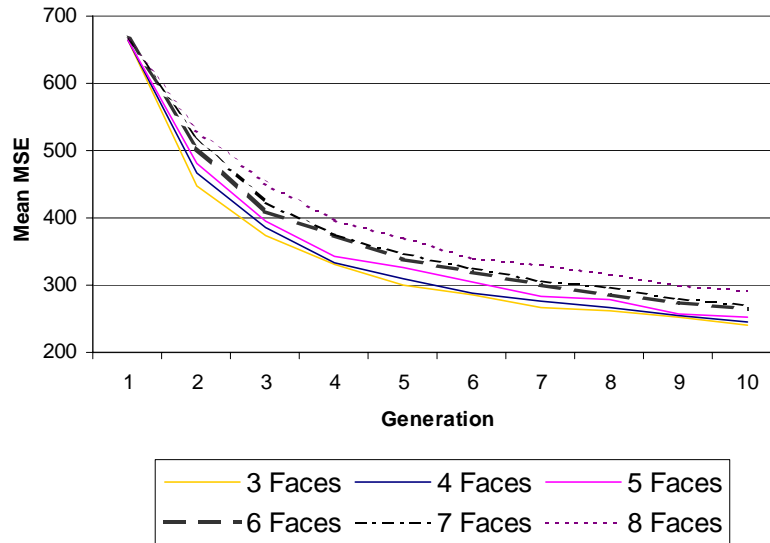
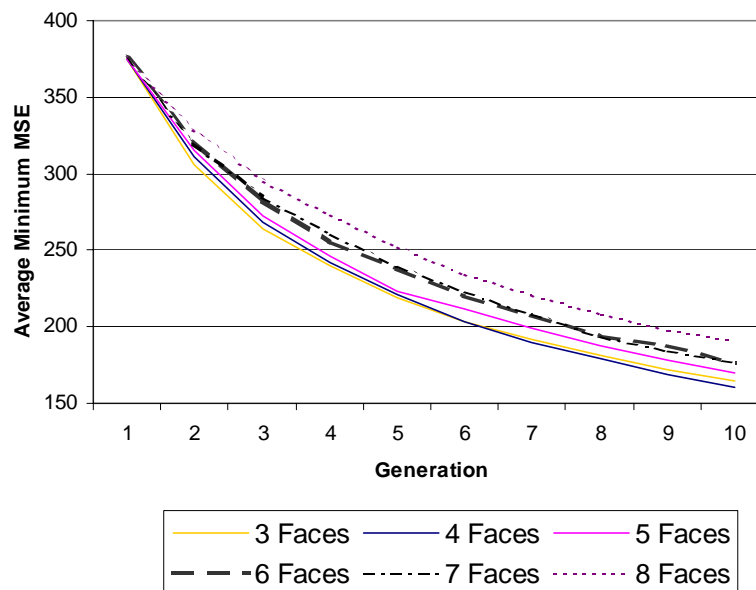


Figure 51: Effect of Varying the Number of Faces Selected on Average Minimum MSE (16 population faces, mutation probability of 0.1, no elitism and 2:1 selection pressure)

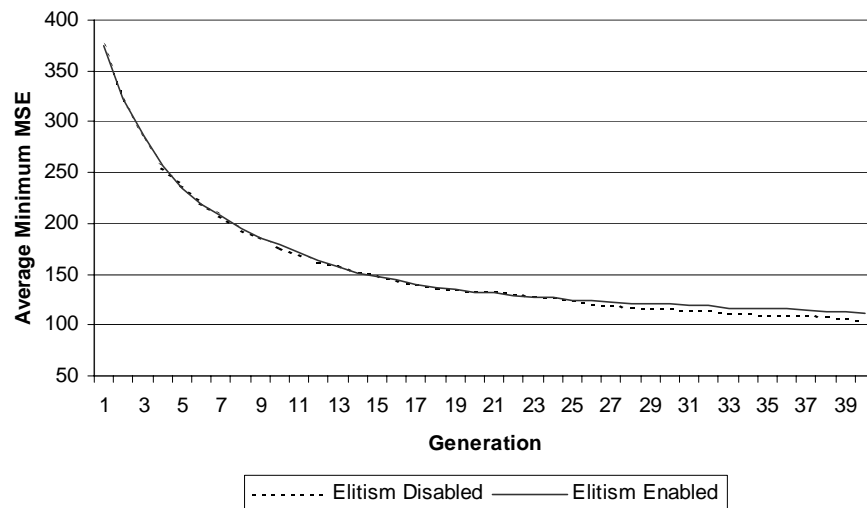


Simulation 5: Elitism

The experiments of Chapter 3 designed a simple elitist mechanism where the best face was always carried forward to the following generation. Such a notion makes intuitive sense as it prevents any “superior” faces from being “lost” through crossover or mutation operators. All the simulations to date have not used this parameter with good reason: the presence of an elitist face will prevent the average minimum MSE scores ever *increasing*, as the lowest MSE will be carried forward, and may mask any undesirable increases⁵³.

Running the simulator with elitism-enabled resulted in an unexpected result for average minimum MSE: when compared against the condition with elitist disabled, the performance is overall worse; referring to Figure 52, this is especially apparent after generation 25. One would expect the most fit individual to be of benefit to evolution. A closer examination revealed a programming error occurred that, although it relocated the image correctly, it failed to correctly copy the relevant shape and texture parameters to the next generation. Subsequent re-selection of that face would result in the incorrect coefficients being used, leading to a decrement in performance.

Figure 52: Initial Effect of Enabling Elitism on Average Minimum MSE (16 population faces, mutation 0.1 and 2:1 selection pressure)



With this mechanism repaired and the simulator re-run, the average minimum MSE appears to be less after the third generation (refer to Figure 53); though this is not statistically reliable over generations 2 to 10 ($t=0.84$, $DF=628$, $p=0.404$). Similarly, the average MSE was

⁵³ As an example of an increasing average minimum MSE, refer to the plot for “0.2” between the 14th and 15th generation in Figure 49.

noticeably less with mutation enabled (Figure 54); this difference approached significance over generations 2 to 10 ($t=1.82$, $DF=628$, $p=0.069$) and was significant over generations 3 to 10 ($t=2.25$, $DF=558$, $p=0.025$). The results indicate that elitism, although not reliably better on locating a best face, can be considered valuable in improving average fitness and should therefore feature in the Face Evolver.

Figure 53: Effect of Repaired Elitism Mechanism on Average Minimum MSE (16 population faces, mutation 0.1 and 2:1 selection pressure)

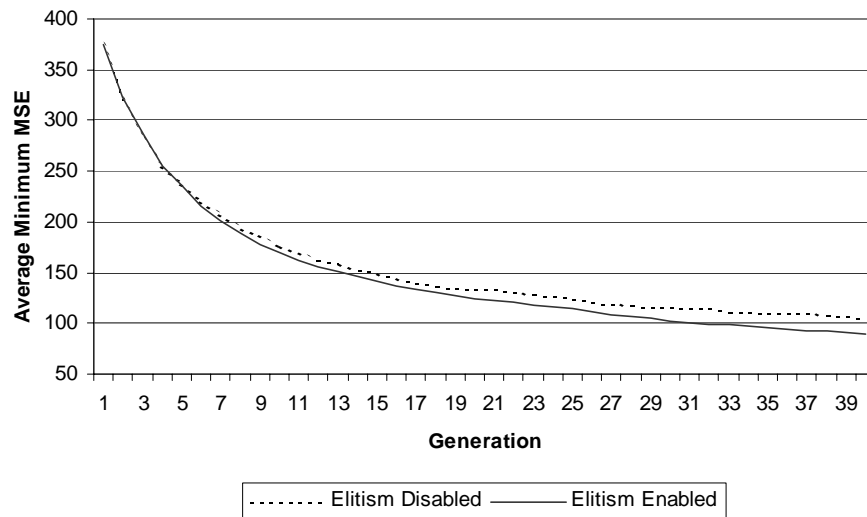
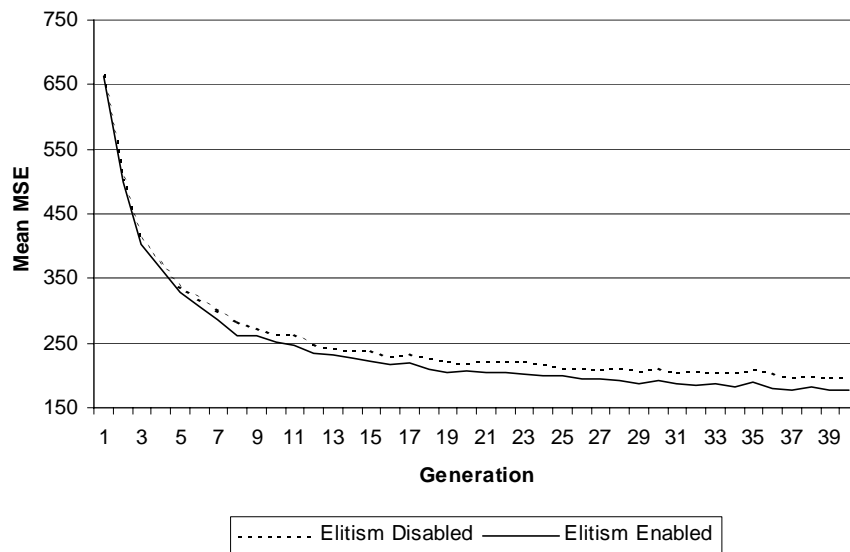


Figure 54: Effect of the Elitism Mechanism on Average MSE (16 population faces, mutation 0.1 and 2:1 selection pressure)



Simulation 6: Combined Effects with Elitism Enabled

It is known that parameters in an evolutionary system tend to interact with each other (e.g. Mitchell, 1996), the consequence of which is that parameters should be evaluated when varied together. Up to this point, the effect of elitism has been largely ignored. It was noted in Simulation 3 that either a mutation rate of 0.05 or 0.1 would appear to be an appropriate setting for this parameter. The effect of elitism has a profound effect though. If simulation is run again with the parameters of Simulation 3 set with a mutation rate of 0.05 and elitism enabled, it is found that the higher mutation rate condition (0.1) results in a *lower* minimum MSE score that approaches significance over the early generations ($t=1.84$, $DF=628$, $p=0.066$). This finding is not unreasonable since higher mutation rates tend to enlarge the area searched by the GA through increases in population diversity. Enlarging the region of search, especially early on in the evolutionary process, can increase the probability of finding a preferable solution. However, higher mutation rates can be too disruptive to performance without the presence of an elitist face. In summary, the higher mutation rate of 0.1 is believed to be a reasonable setting for the Face Evolver especially when used in conjunction with elitism.

On the other hand, it was proposed in Simulation 2 that the selection pressure mechanism on the best face be set to unity (rather than 2:1 or even 3:1) since the minimum MSE measure was significantly lower with this setting. However, re-running the simulation for elitism enabled (0.1 mutation rate) with a 2:1 selection pressure results in a *lower* minimum MSE compared with a unity selection pressure; although the effect size is small (4.3) and non significant ($t=1.00$, $DF=628$, $p=0.313$). Overall, it was not considered detrimental then to leave this parameter setting at 2:1 in the Face Evolver.

Simulation 7: Population Size (Revisited)

Now that a set of parameters have been proposed for mutation (0.1), selection pressure (2:1) and elitism (enabled), the investigation returns to the appropriate number of faces required in a population. The important finding from Experiment 7, Experiment 8 and Simulation 1 was that an increase in population size tended to result in a higher rate of decrease in the mean MSE and a lower overall terminal value. One may ask therefore whether there is any benefit in employing a larger population for fewer evolutionary cycles? It was estimated in Chapter 3 that 10 generations of 16 faces is likely to be sufficient to produce a face of “acceptable” likeness; the generation of 160 faces. Could this number of faces be better distributed then over fewer cycles with a larger population?

This notion can be explored in simulations that study performance when the number of generations is manipulated, with the number of population faces adjusted accordingly, so that 160 faces are always generated. Ultimately, this was tested by increasing the number of

generations (N_g) from 1 to 24. In each case, the number of faces in a population (N_p) was computed by the equation –

$$N_p = 160 / N_g \quad \dots (2)$$

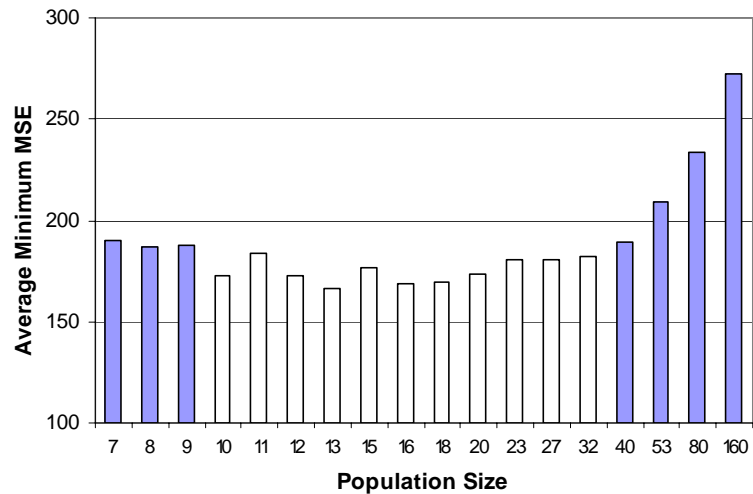
In simulations, the equation was rounded to the nearest integer. For example, with 7 generations, a population size of 23 (22.9) faces would be required. This necessarily leads to error, the maximum occurring for a population size of 15, with 5 more faces than the average (160) being produced. This error (3.1%) is considered too small to be of significance.

The simulator was then run.

The data in Figure 55 was calculated by averaging the minimum MSE taken from the last generation for each target. It can be seen from the graph that the average minimum MSE appears roughly equivalent in the range of 10 to 20 population faces. Outside this range, the error becomes worse, especially for the relatively larger population sizes. T-tests revealed no significance difference in the average minimum MSE in the range 10 to 32 population faces compared with a population size of 16 ($p > 0.05$).

This should mean that if the number of faces presented is taken into account, then the benefit of doubling the population size found in Chapter 3 should be eliminated. This can be easily verified by comparing the rating scores from generations 1 to 3 for a population size of 32 (Experiment 7 and Experiment 8) with those from generation 2, 4 and 6 for a population size of 16 (Experiment 2 and Experiment 4). When this analysis was run, a two factor ANOVA was significant for familiarity (famous or randomly-generated targets; $F=25.0$, $DF=(1,185)$, $p < 0.001$), but not for population size (16 or 32 population faces; $F=0.01$, $DF=(1,185)$, $p < 0.927$) and there was no interaction ($F=0.71$, $DF=(1,185)$, $p < 0.401$). As seen before, the famous targets were evolved better than randomly generated ones, but as predicted, when the number of faces presented was kept constant, there is no longer a benefit for an increase in population size. The result is therefore consistent with the prediction arising from simulation.

Figure 55: Effect of Varying the Population Size on Average Minimum MSE (mutation 0.1, 2:1 selection pressure, 6 faces selected and elitism enabled); the white bars indicate a non-significant difference compared with a population size of 16 faces.



Comparing the average minimum MSE (Figure 55) for a population size of 16 and 160 does illustrate the benefit enjoyed by evolution (a 38% decrease in MSE). It is interesting to note that even a single cycle of evolution (i.e. 80 faces bred together once) results in a large decrease in MSE (14%). Note also that although evolution is of value, the change from a “flat” evolutionary system (i.e. Simulation 1: no mutation, no elitism and unity selection pressure) to an “optimized” one (a mutation rate of 0.1 with elitism and a 2:1 selection pressure) does not represent a huge reduction (14%) in the average minimum measure; the difference is nonetheless significant ($t=7.76$, $DF=768$, $p=0.006$).

Overall, this result suggests that photofits are best produced by evolution (rather than the random generation of faces). In addition, at least 5 generations (with 32 faces) would appear necessary to achieve good performance; more evolutionary cycles with relatively smaller populations seem to make little difference (down to a minimum of 10 faces and 16 generations). In conclusion, the currently adopted value of 16 population faces would seem to be appropriate, although there is no evidence for significant differences over the range 10 to 32 faces.

The following simulations explore two more aspects of face evolution and suggest ways to further increase performance.

Simulation 8: Coefficient Pruning

It is clear from Simulation 1 that increasing the number of population faces, with the number of generations kept constant, results in better performance. However, if many of the faces have a high degree of similarity, then the effective size of a population will be reduced. What might be desirable then would be to identify and remove those faces that are very similar

to each other. Of course, this process will necessarily reduce the number of population faces. Therefore, it appears sensible to pre-generate more faces than are ultimately required and to eliminate the most similar ones.

One method to decide which faces should be eliminated is to compute a distance error score (e.g. MSE) between all possible combinations of faces and then to remove those with the closest error. As this is likely to be time consuming, especially with a large number of starting faces, a simple iterative method is proposed: rather than computing all possible combinations of error score, one computes the error scores from one of the faces chosen at random. The face that has the lowest error score is then removed or “pruned” from the population. The process is then repeated until the desired population size is reached.

One problem is that it is computationally expensive to morph a face (about 300ms) and it could result in excessive delays in creating a population. To overcome this difficulty, it was decided to eliminate similar faces based on face coefficients rather than their reconstructed representations. Hence, an excess of shape and texture coefficients would be produced and the most similar ones removed. It appeared sensible to examine the “pruning” of texture and shape coefficients separately; pruning could be performed on populations that were twice and three times larger than required. Other system parameters were set as in Simulation 6 (with elitism enabled).

For the shape coefficients (Figure 56), the result of pruning on the minimum MSE can be seen to be better for a 2:1 prune compared with either a 3:1 prune or the baseline condition (no prune) up to about generation 10. Beyond this point, the baseline condition appears better. Indeed, there is a significant reduction in the average minimum MSE for the 2:1 prune condition over generations 2 to 10 compared with the baseline ($t=3.31$, $DF=628$, $p=0.001$). Similarly, for the texture coefficients (Figure 57), there appears facilitation in performance up to generation 5 (not as marked as for shape) but fails to reach significance ($t=1.05$, $DF=278$, $p=0.295$). It is proposed that a 2:1 pruning mechanism be implemented for the face shape components of the Face Evolver, but left optional for texture.

Figure 56: Effect of Shape Pruning on Average Minimum MSE (16 population faces, mutation 0.1, elitism enabled and 2:1 selection pressure)

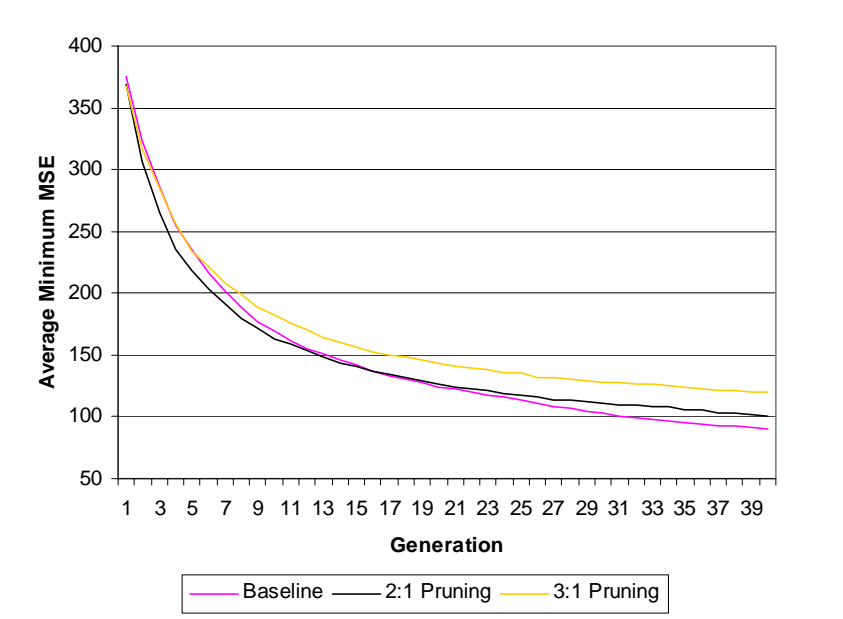
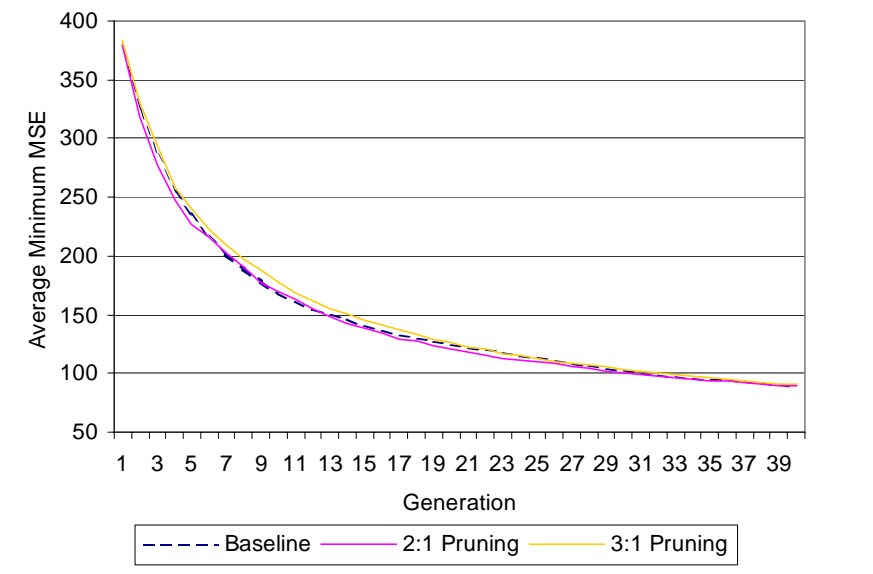


Figure 57: Effect of Texture Pruning on Average Minimum MSE (16 population faces, mutation 0.1, elitism enabled and 2:1 selection pressure)

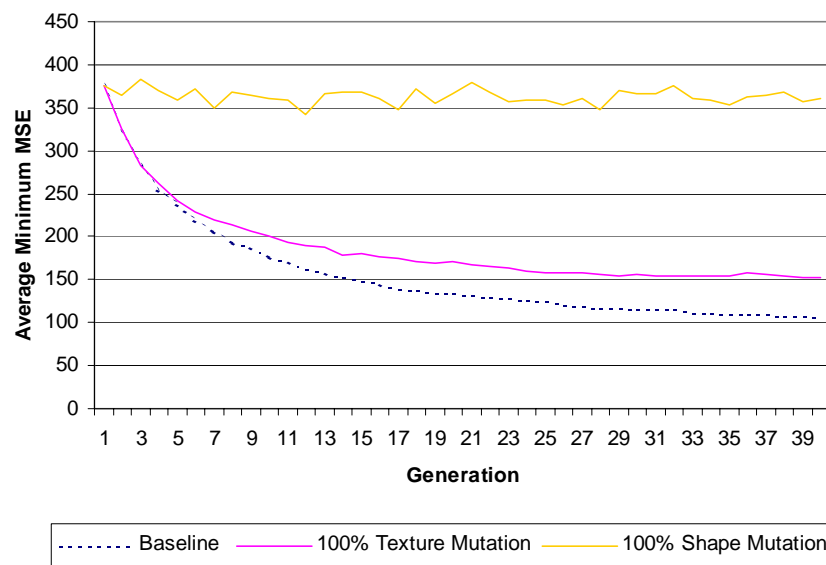


Simulation 9: Separate Selection Mechanisms

In situations where the probability of mutation is set to the maximum of 1.0, no evolution can occur as random faces are produced (i.e. shape and texture coefficients always take on random values). A curious result was observed when comparing simulations run with the mutation rate set to 1.0, first for the shape coefficients and then for the texture coefficients;

run with no coefficient pruning, no elitism and the remaining parameters of Simulation 8. A plot of the average minimum MSE (Figure 58) reveals that there is no benefit of evolution with 100% shape mutation. In contrast, there appears little difference in the performance to the baseline with 100% texture evolution over the first 3 generations, then performance is noticeably worse. Taken together, this indicates that (1) evolution of texture is dependent on evolution of shape (and not vice versa) and (2) shape is more important initially with texture playing an increasingly more important role later.

Figure 58: Effect of the Maximum Mutation Rate (probability of 1.0) on Average Minimum MSE (16 population faces, mutation 0.1, elitism off and 2:1 selection pressure)



Clearly, shape evolution results in a tendency for pixels in a population to become coincident with those of the target. In other words, features tend to become *aligned*. The alignment of features is likely to result in lower error scores, especially for the overall head shape and hairstyle due to the relatively large proportion of the image occupied by them (Bruce & Young, 1986). It was noted that 32 (out of 35) of the targets have qualitatively darker hair anyway and will result in proportionally high error scores for mis-alignment in this feature than with more average intensity styles (such as mid brown). It would appear sensible then for the shape and texture components of a face to be treated more independently during evolution. One can easily imagine a situation where one of the population faces has a relatively good texture but a poor shape; the current implementation would tend not to select such a candidate (as the error from the shape is too high). A simple solution appears to be to select the texture components independently of shape components.

This hypothesis can be tested in simulation by choosing face *shapes* with the lowest MSE with respect to shape, and face *textures* with the lowest MSE with respect to texture. As

previous simulations have tended to select 6 faces per population, it is proposed that the same number of shape and texture components be used (6 sets of texture components and 6 sets of shape components). The only complicating factor is how to select the best face overall. The simplest solution appears to be to use the existing method of choosing the population face that has overall the lowest MSE.

When run again, with just the method of face selection changed, only 22% of the time was the same face selected (for both shape and texture); an average of 1.3 faces per 6 selected. In contrast, the overall best face was found to be selected for both shape and texture 70% of the time among those selected for both. The effect on the average MSE is minimal (Figure 59), and non significant ($t=0.71$, $DF=628$, $p=0.480$) over generations 2 to 10, but is far more pronounced for the average minimum MSE (Figure 60) and approaches significance over the same range ($t=1.80$, $DF=628$, $p=0.073$). The poor overlap of jointly selected faces (22%) together with an approaching significant improvement on the minimum MSE measure suggests that separate mechanisms for evolution would be a worthwhile development to the Face Evolver.

Figure 59: Effect of Separate Evolution for Shape and Texture Components on Average MSE (16 population faces, mutation 0.1, elitism off and 2:1 selection pressure)

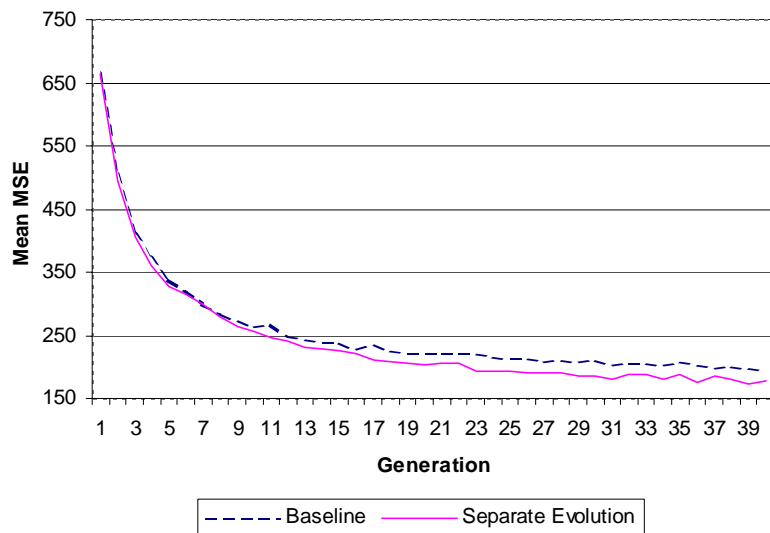
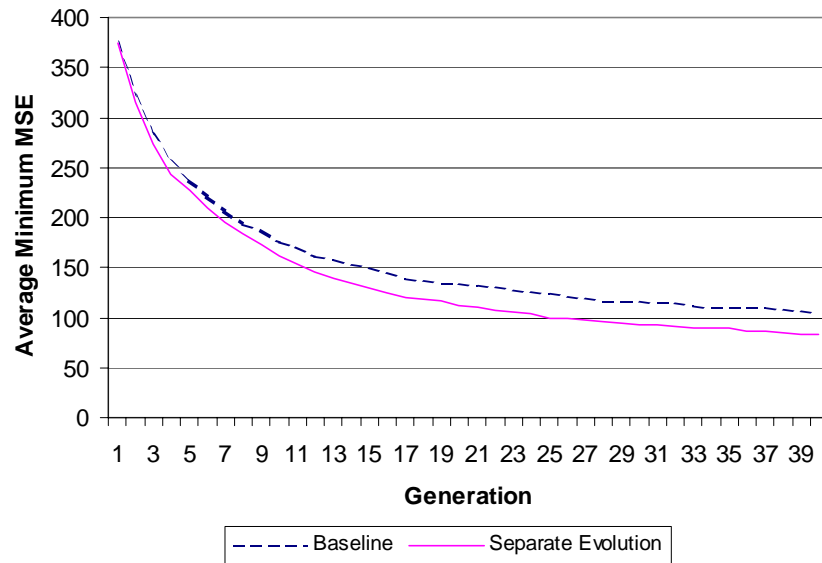


Figure 60: Effect of Separate Evolution for Shape and Texture Components on Average Minimum MSE (16 population faces, mutation 0.1, elitism off and 2:1 selection pressure)



General Discussion

The simulations in this chapter have been instrumental in beginning to understand the role of parameter settings in the Face Evolver. It was found that settings of 0.05 and 0.1 were appropriate for mutation. Elitism was also found to be of value - both in itself and also enabling the higher mutation rate of 0.1 to be preferable to the 0.05 setting. The selection pressure was initially proposed kept constant for all selected faces but, in the presence of elitism, was found to be slightly beneficial at the 2:1 value used in Chapter 3.

Initially, the population size was also found to be positively related to the average and peak system performance, with the error scores of larger populations appearing to decrease faster and to a lower terminal value than smaller populations. However, when the number of faces generated by the Face Evolver was kept constant (160 faces), it was found that there were no significant differences in peak system performance (via the minimum MSE) for population sizes over the range of 12 to 32 faces. This indicates that 16 population faces used for the Face Evolver is not an inappropriate size. Indeed, a re-analysis of the evolution data from Chapter 3 revealed that once the number of faces generated was controlled, there was no longer any benefit for increases in population size.

It was also found that the clustering of faces was important. A process of two-fold over-generation and the successive selection and pruning of faces that were most similar resulted in significantly better peak performance for the shape components and, to a lesser degree, the texture components.

Another area indicating change to the current approach was in the number of faces selected from a population. The data clearly indicates the benefit to the average and peak performance for selecting fewer faces. In fact, just reducing the number to 5 faces made a significant impact on the minimum error scores; the lower limit was suggested to be either 3 or 4 faces.

Finally, a further simulation established that it was valuable to implement separate selection mechanisms for shape and texture. Anecdotally, from an operator's perspective, it was clearly apparent when a composite was being created that sometimes a population face would be generated with a relatively good *texture* (to a target) but not a good *shape* (and vice versa). When this happened, a witness would be forced to grudgingly accept a poor quality representation or make another choice. Separating shape from texture not only avoids dissonance, but also acknowledges the role of featural and configural information in face perception.

Of course, both the features of a face and their configuration are modeled separately in EvoFIT: while *shape* models configural changes, *texture* models featural changes. It is clear that the information about features plus the information about the relationships between features are important in face perception (e.g. Bruce & Young, 1999). Interestingly, Cabeza & Kato (2000) suggest that features and configuration have equal salience. Their work involved *prototype* faces. One type, a "featural" prototype, contains facial features taken from the different faces seen at study. Research has frequently demonstrated that subjects tend to misidentify these prototypes at test, even though they have not been seen in the composite form previously (e.g. Cabeza, Bruce, Kato & Oda, 1999; Inn, Walden & Solso, 1993; & Solso & McCarthy, 1981). Cabeza & Kato demonstrated that the false alarm rate for a featural prototype was the same as a configural prototype (a face with the same configuration as the test set), indicating equivalent importance for features and their configuration in face perception.

Parameter settings

A summary of suggested parameter values is shown in the table below –

<u>Parameter</u>	<u>Setting</u>
Population Size (Np)	10-32
Generations (Ng)	5-12 (depending on population size)
Mutation Rate	0.1
Selection Pressure	2:1
Elitism	✓
Component Pruning	2:1 Shape and (optionally) 2:1 Texture
Selected Faces	Fewer is better (lower limit 3 or 4)
Separate Selection of Shapes and Textures	✓

Table 6: Summary of Suggested Parameter Values (values in bold indicate new settings and features)

The parameters with a bold highlighting in Table 6 indicate settings that could be changed (i.e. population size, number of generations, mutation rate and the number of selected faces) or where new features could be added (i.e. pruning and the separate selection of face shapes and textures). Note that all the parameters listed in the table except the last two concerns the generation of faces. The latter two, the number of selected faces and the separate selection of shapes and textures, are issues more closely associated with system usability. The following chapter develops the Face Evolver with these changes in mind as well as the ones proposed in the previous chapter (e.g. hairstyle modification). The result is an implementation that can be thought of as a *photofit system*: EvoFIT.

Chapter 5: The EvoFIT System

The objective of this chapter is to implement the changes proposed in the previous two chapters of this thesis. In addition to these points, further design features are added, particularly those likely to promote rapid convergence to a target. These include the availability of four more face palettes and a utility that can directly manipulate facial features. The result is a photofit system: EvoFIT.

During the evaluation of this system, encouraging results were found using EvoFITs created from the memory of unfamiliar faces. A later evaluation involved famous faces also constructed from memory and demonstrated a spontaneous naming rate of about 10%. This was found to be about 7% less than EFITs constructed under the same conditions. Evidence is presented demonstrating that celebrity age was a factor preventing better EvoFIT performance. It was shown that around 30 years is a likely upper age limit for the implemented EvoFIT database. A follow-up study found spontaneous naming rates at 25% for novice operators working with the target present during the construction of more appropriately aged famous faces. This study also demonstrates that facial distinctiveness is expressed in photofits composed with EvoFIT. Interesting results are also presented where EvoFIT was used in a real case.

What Must Be Done?

Analysis of data from Chapter 3 indicated that if photofitting is carried out from targets obtained external to the Face Evolver, a necessary condition for real life situations, then improvements need be made to allow the use of better hairstyles as well as to provide more complex shape and texture models. Chapter 4 recommended that the parameter settings for mutation (0.1), selection pressure (2:1) and elitism should be adopted. It was also proposed that, for faster convergence to a solution, shape and texture coefficient pruning should be used, the number of selected faces should be limited as far as possible (to a minimum of 3 faces) and the selection procedure should be made on the basis of separate shape and texture information.

Multiple Face Palettes

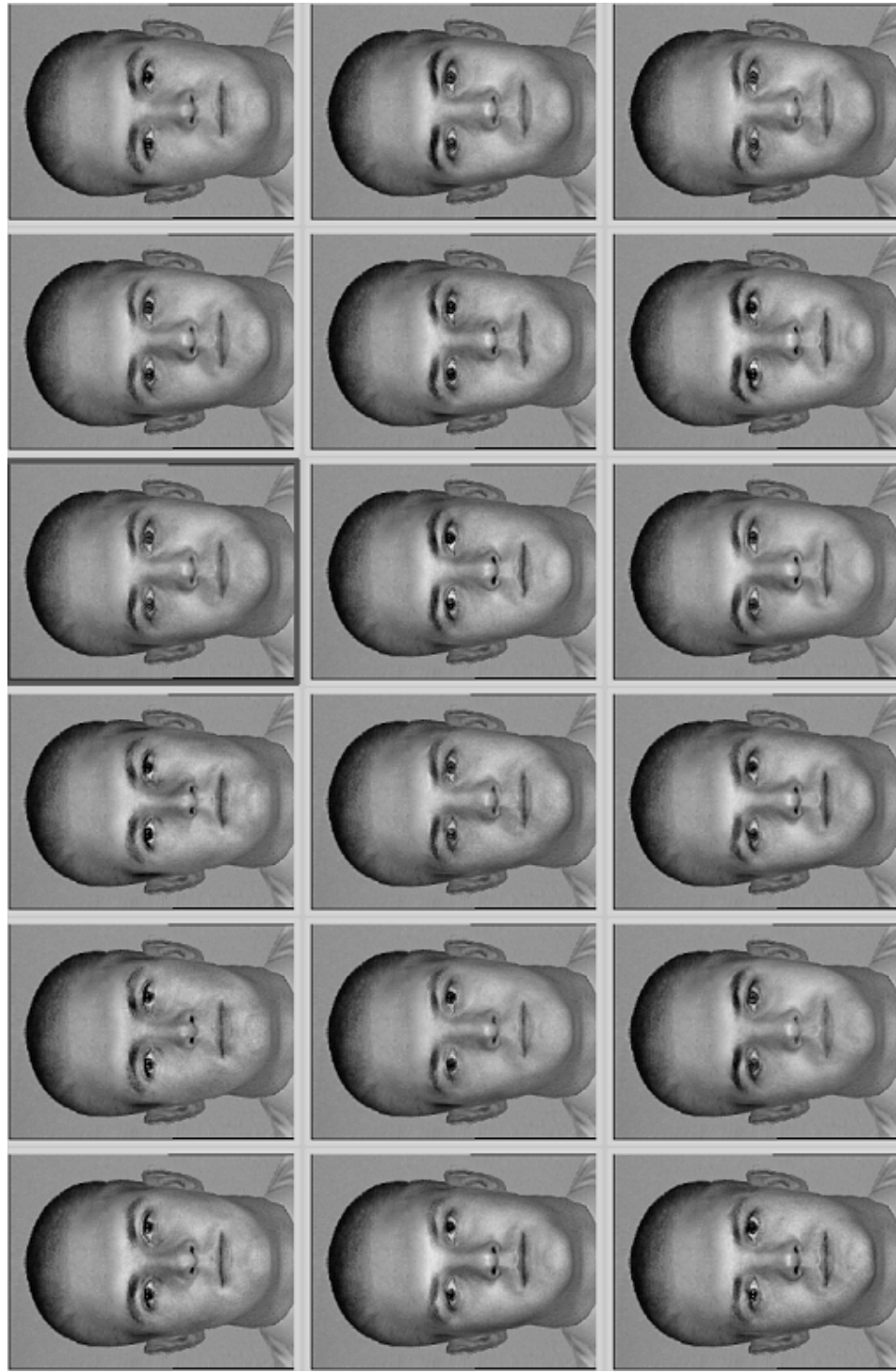
Intuitively, it is quite natural for shape and texture information to be selected separately: there is of course a separate model implemented for each and there is no immediate reason to suppose that information should be related between the two models. One would not expect within a Caucasian database, for example, that the colour of a mouth would be related to its size or spatial location. But, how should separate texture and shape information be presented to a user? In much the same way as faces are displayed currently, a simple design would involve the presentation of one screen or “palette” for facial shape, and another for facial texture; referred to hereafter as the Facial Shape Palette (FSP) and the Facial Texture Palette (FTP) respectively.

A representation potentially relevant for the texture palette (FTP) is already available. Recall in the generation of a face that a “shape-free” (texture) image is initially produced from the texture model (prior to a morph with a shape vector defined by the shape model). This texture representation could be used on the FTP. In contrast, to produce an FSP, one could create a set of “wire-frames” based on the shape vectors for each face. There is however evidence from face recognition studies that a drop off in performance can occur when an image is reduced to a line drawing (Bruce, Hanna, Dench, Healey & Burton, 1992; Davies, 1982, 1983b; Davies, Ellis & Shepherd, 1978; Leder, 1996; Perrett, Benson, Hietanen, Oram & Dittrich, 1995; and Rhodes, Brennan & Carey, 1987) and this suggests that another representation may be preferable.

It was thought that the average texture (of the database) could be used as a basis on which to produce a shape morph. Recall that Experiment 4 and Experiment 8 (in Chapter 3) contained a palette where each hairstyle had been superimposed onto the average texture of the database. As there appeared to be no reported problems with this representation for the selection of hairstyle⁵⁴, the average texture was used. Facial shapes could then be produced by a morph defined by a face’s shape vector on the average texture. An example of an FSP and an FTP is shown in Figure 61 and Figure 62 overleaf –

⁵⁴ This is based on personal discussion with participants coupled with a lack of comments (left by participants) regarding difficulties in the selection of hair.

Figure 62: An Example Facial Texture Palette (FTP) Displaying a set of Randomly Generated Textures



The overall result is that there are now 3 representations for each face: one for shape, one for texture and one displaying both shape and texture. The combined representation of shape and texture is the same as that used in Chapter 3 and is now referred to as the Facial Normal Palette (FNP). The FNP would have the function of highlighting those faces selected for shape and texture. A simple colour-coding scheme was adopted: a blue border would be assigned to faces selected for shape, a green border for texture, and a red border for both shape and texture. As before, provision could be made to de-select any unwanted shape and/or texture selections (on any of the palettes).

It is still required though for the user to select a face that is perceptually the closest (i.e. a best face). As separate selections are now to be made for shape and texture, it appears logical that the assignation of the best face be carried out on the FNP after visiting the FSP and FTP. Note that this procedure now differs from the Mark II system, where the best face was made *before* other selections. This procedure has the advantage that it reduces the search complexity by selecting the best face from only the most similar faces rather than an entire population. This procedure is similar to Baker & Seltzer (1998) who asked subjects initially to select 5 faces (that were most similar to an assailant from a set of 100 images), before putting them into rank order.

Experiment 9: Separate Shape and Texture Selection

As this modified selection procedure was rather radical, it was decided to run a small pilot experiment to test the effectiveness of the new interface. To achieve this, a total of 9 targets were randomly generated and a palette made available for hairstyle selection (as in Experiment 4 and Experiment 8). Three subjects first created a photofit with the target on display in the centre of the screen (Condition A), then with a second target available for *inspection*⁵⁵ (Condition B) and lastly with a further target created after a 1 minute exposure (Condition C). A minimum of 2 generations were allowed and the session was terminated when subjects judged that an acceptable likeness had been reached. The “perceptually-closest” face was then saved on disk as “the photofit”.

Subjects were instructed that they needed to select at least 4 faces from both the Shapes (FSP) and Textures (FTP) palettes. The remaining parameters were largely the same as Experiment 1 (except 16 population faces were used for Condition A and 18 for Conditions B and C⁵⁶; selection pressure was set at 2:1; mutation rate was set at 0.1 and the [repaired] elitism mechanism was enabled). All photofits were constructed in the presence of an operator [me]

⁵⁵ In the inspection condition, the target could not be seen at the same time as the population faces but could be referred to as often as required. This condition is considered more challenging than having the target on display but not so hard as a one-shot memory condition.

who was responsible for controlling the software. The operator did not see the targets for Conditions B and C, nor did he offer any suggestion regarding the selection of faces for a given target. The following 9 EvoFITs were created -

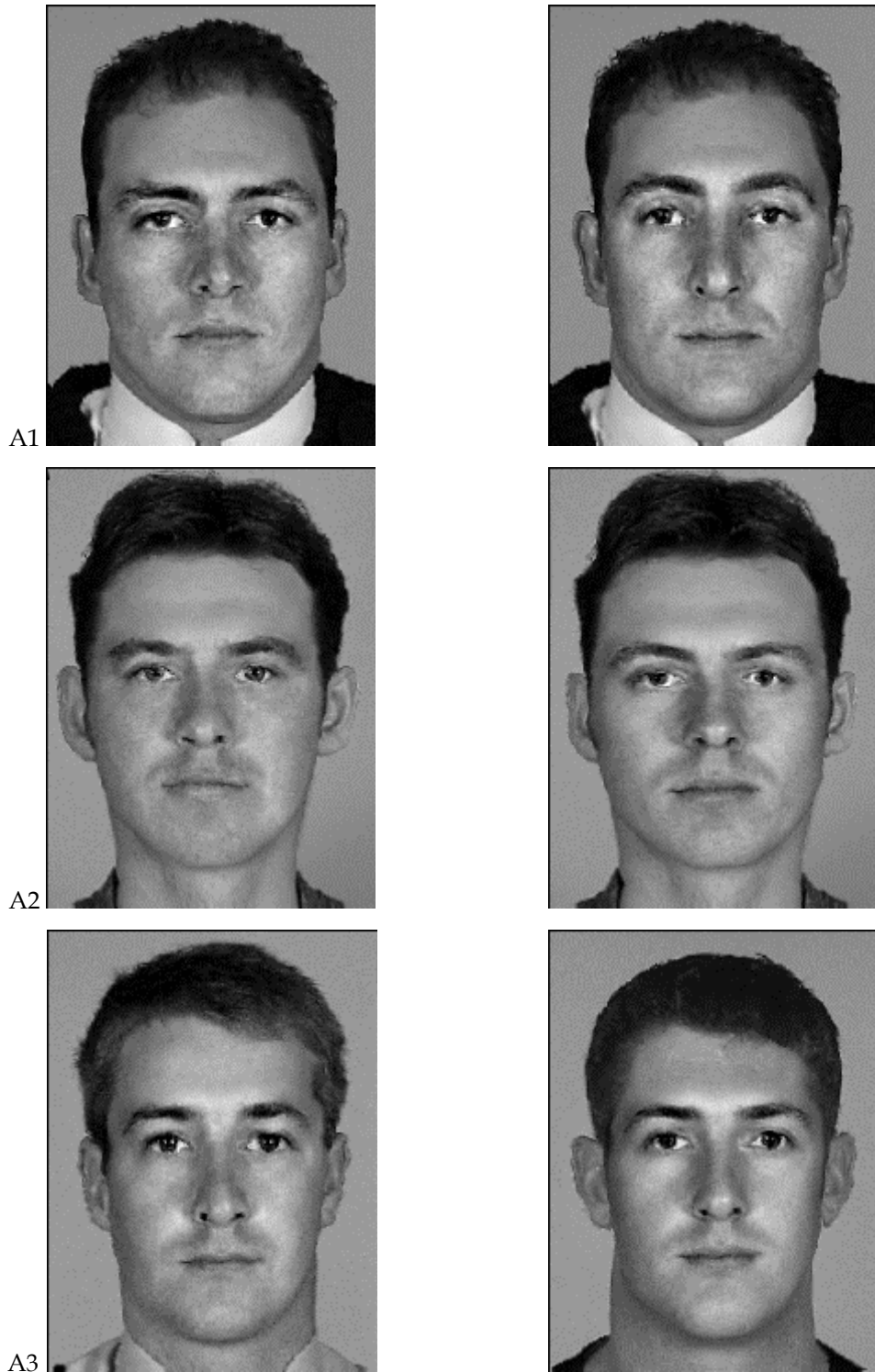
⁵⁶ As no target is displayed in the centre of the screen, a total of 18 population faces can now be viewed.

Figure 63: EvoFITs Created with Target Visible (A), from Inspection (B) and from Memory (C)

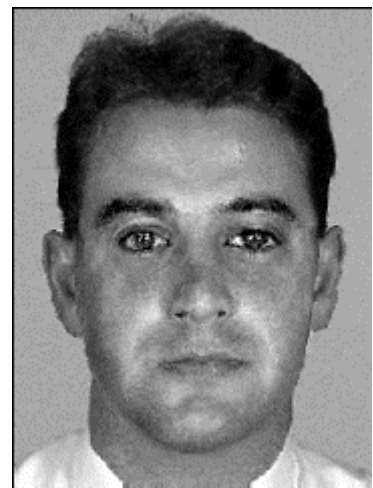
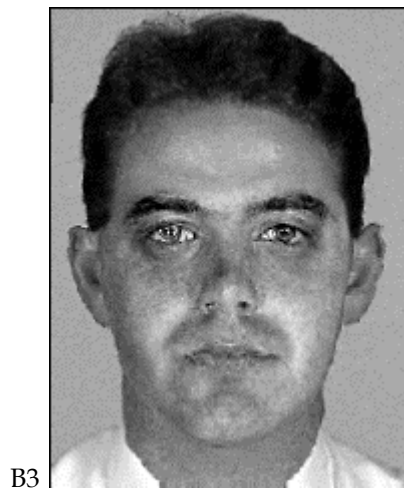
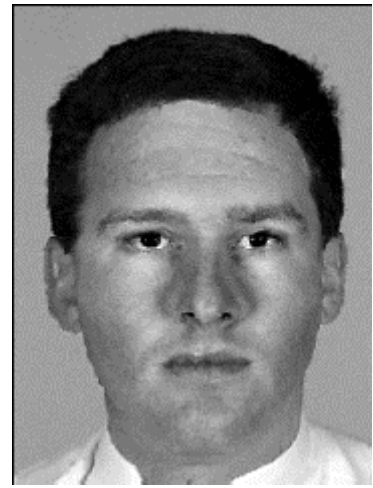
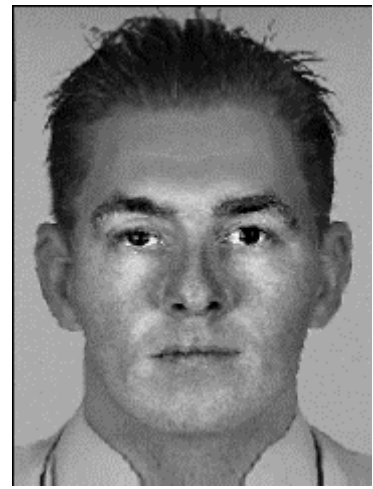
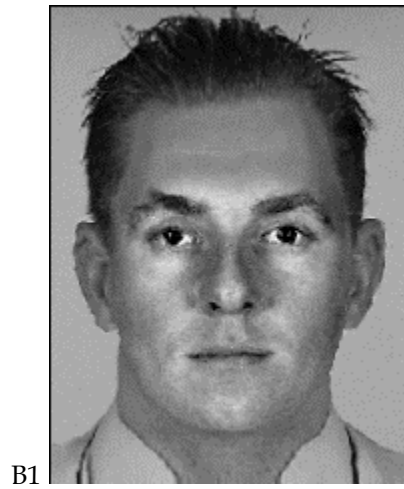
(A) EvoFITs Created from Target Visible

Target

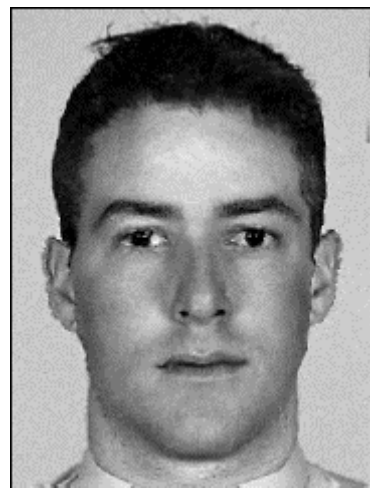
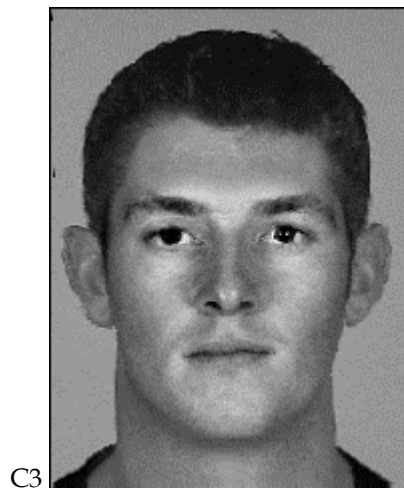
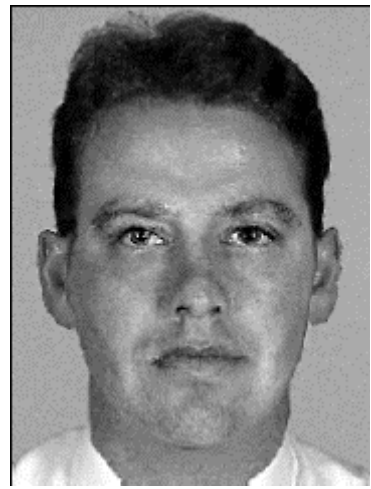
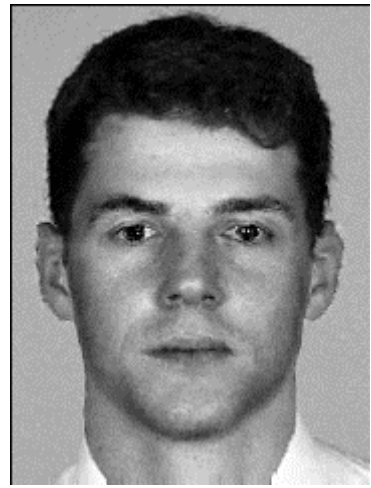
Best (EvoFIT)



(B) EvoFITs Created from Inspection



(C) EvoFITs Created from Memory



Evaluation was carried out by showing all 9 photofits to a different set of participants (35 in total) and asking them for a rating on the AFSS (refer to Table 2 in Chapter 3) in the presence of the relevant target. The order of the presentation was randomized for each subject.

Results

The overall average rating was 6.7. The average rating was 6.3 for Condition A, 7.7 for Condition B and 5.3 for Condition C. A repeated-measures ANOVA did not find a significant effect for condition ($F=2.94$, $DF=(2, 6)$, $p=0.129$). Examination of the photofits revealed that the “correct” hairstyle (i.e. not the same hairpiece as appearing on the target) was not selected for EvoFITs A3, C1 and C3 (Figure 63). The average rating was 7.5 for photofits with the same hairstyle and 5.0 without. An ANOVA reveals that photofits with the correct hairstyle were rated significantly higher than those without ($F=26.77$, $DF=(1, 4)$, $p=0.007$); there was still no main effect for condition ($F=1.07$, $DF=(2, 4)$, $p=0.424$) and there was no evidence of an interaction ($F=0.82$, $DF=(1, 4)$, $p=0.417$). Considering data from photofits with the correct hairstyle selected, the average rating was 7.3 for Condition A, 7.7 for Condition B and 7.2 for Condition C.

Discussion

The overall rating (6.7) does show a good degree of satisfaction with the photofits in general: a rating in the semantic category of “Many similarities”. The selection of the exact hairstyle resulted in a large and highly significant 2.5 point increase on the AFSS, demonstrating once again the importance of this feature. Interestingly, there were proportionally more correct hairstyles (5/6) in the non-memory conditions (Conditions A and B) compared with the memory (1/3) condition (Condition C), suggesting a detrimental effect of memory on hairstyle selection⁵⁷. In this small study, it is concluded that a high degree of satisfaction can be achieved from the photofits created using the new selection technique. It also indicates how memory can play an important role in choosing a hairstyle and that rating scores can be strongly modulated by this chosen hairstyle.

More Palettes?

During Experiment 9, it was observed that the selection of shapes on the FSP and the selection of textures on the FTP became increasingly more difficult. It was the case that some participants did not wish to visit these screens, opting for the FNP instead. This occurred because of the increasing disparity between the average texture used for the FSP and the

⁵⁷ Note that strong conclusions are deliberately not made regarding hairstyle here due to the small number of photofits created.

target's texture as the population became more like the target with evolution (and similarly for the average shape used for the FTP and the target's shape).

An improvement appeared to be to somehow "adapt" the average texture to make it more like the target with increasing generation (and similarly for the average shape used for the FTP). Arguably, the closest representation to a target at anytime is the best face that was selected in the previous generation. It is proposed that the shape and texture components of this face are used as a basis for the FTP and FSP respectively. It is hypothesized that this mechanism should enable the FSP and FTP to be more representative of the target. Two more face palettes need be added to achieve this; these are referred to as the FSPBT (Facial Shape Palette with Best Texture) and the FTPST (Facial Texture Palette with Best Shape). It is intended that these palettes be evaluated by user feedback during the next evaluation session. An example of each palette can be seen on the following 2 pages -

Figure 64: An Example Facial Shape Palette with Best Texture (FSPBT) for a set of Randomly Generated Shapes

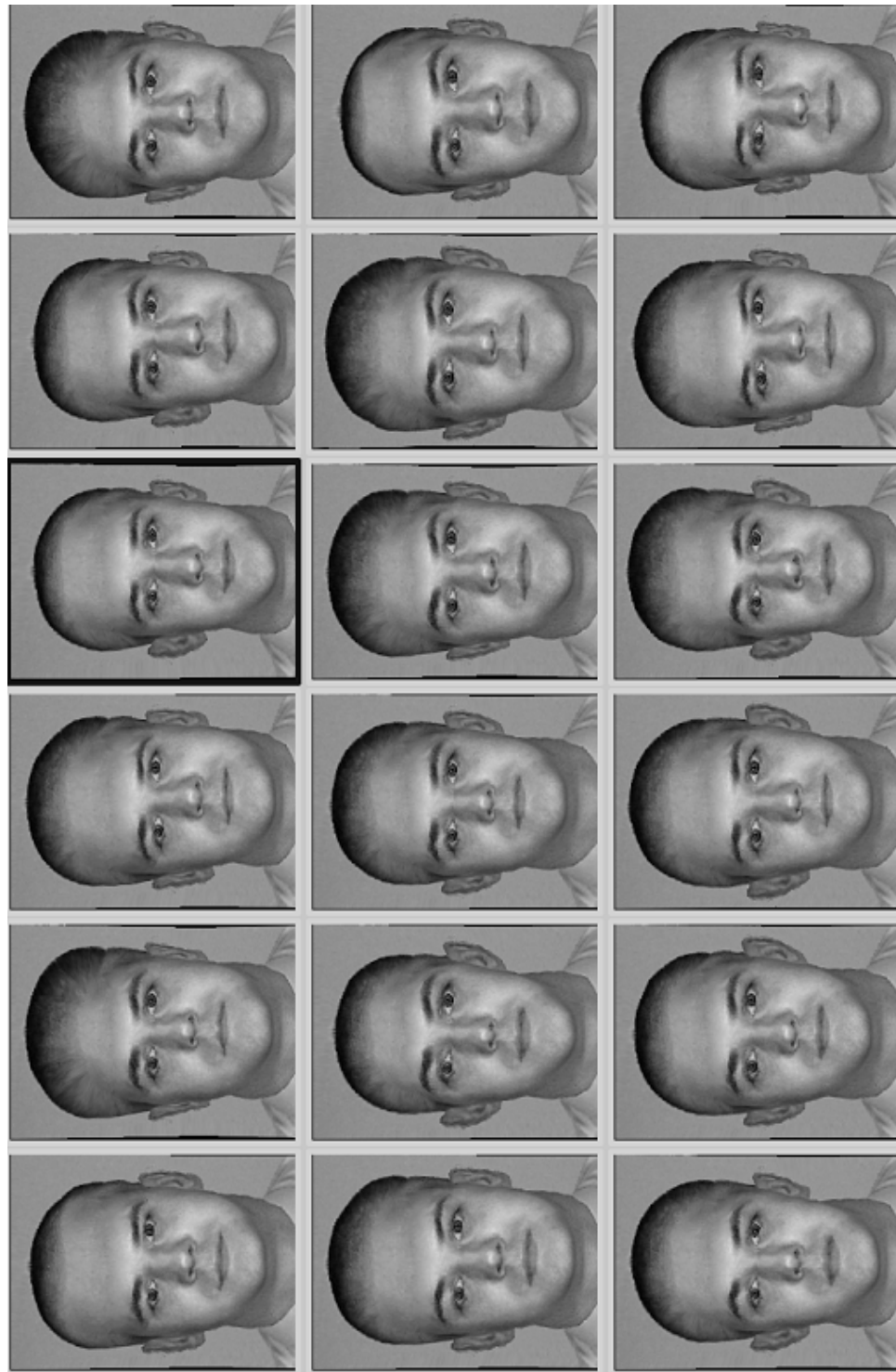
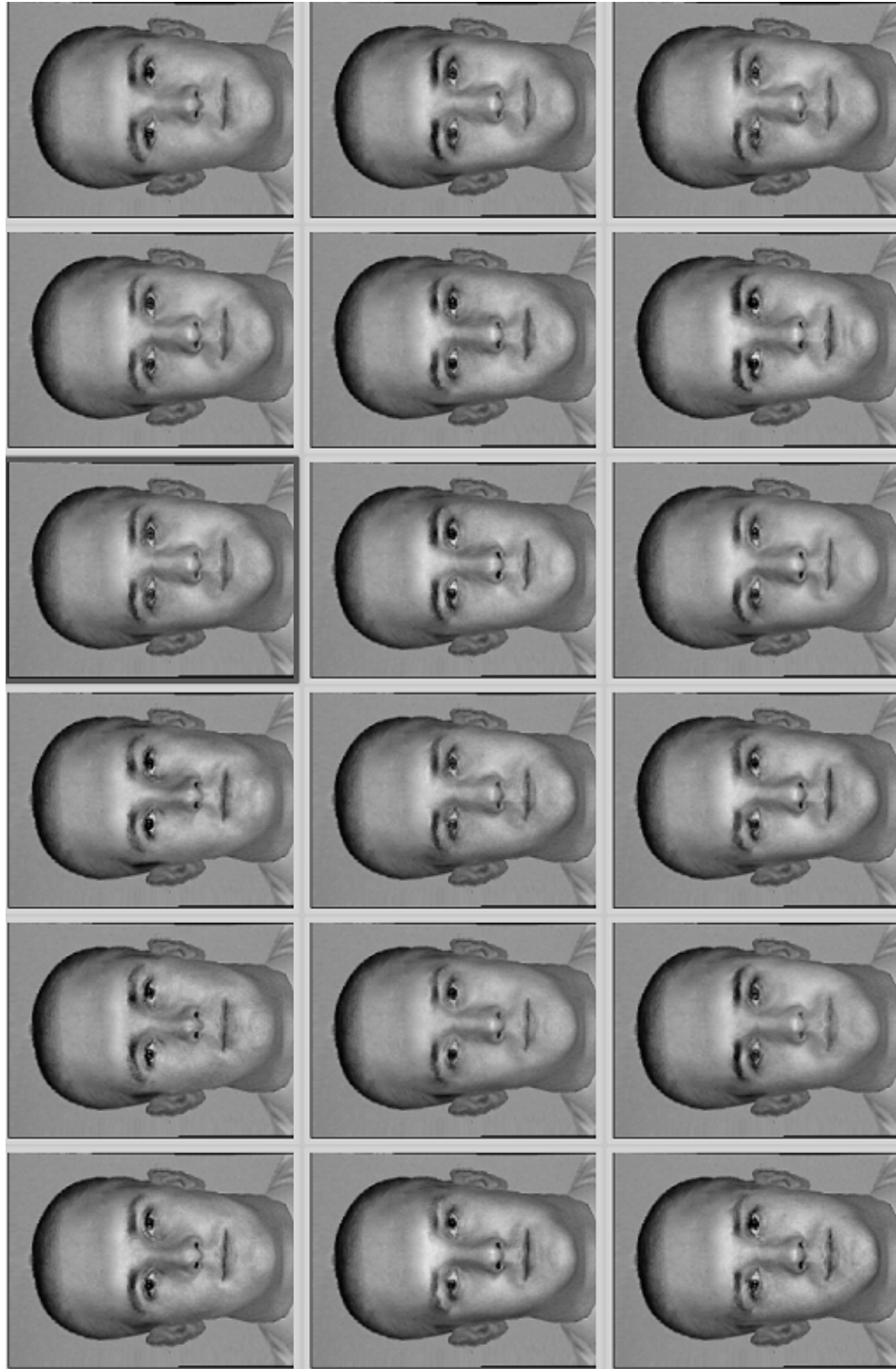


Figure 65: An Example Facial Texture Palette with Best Shape (FTPBS) for a set of Randomly Generated Textures



During the development of the FSPBT and FTPBS, it was envisioned that a face appearing in these palettes might sometimes be a particularly good face; i.e. a face better than the best in the previous generation. Unfortunately, switching to the FNP would lose such a potentially advantageous representation. This seemed an undesirable feature (and might cause great distress to a witness who had arrived at a preferable likeness). A simple solution was to add a system function that enabled desirable faces on the FSPBT and the FTPBS to be included in the FNP.

Another development considered likely to increase evolutionary efficiency, also raised during photofitting sessions, was the ability to combine information between faces on the FNP. Several subjects commented that they believed the shape on one face was “good” and the texture on another was “good”. The addition of another function, known as the Facial Composite Tool, combining the shape from one face and the texture from another was implemented for the EvoFIT system as well.

Increasing the Complexity of the Face Model

It was proposed in Chapter 3 that the number of faces used to construct the shape and texture models be increased from the current database size of 35. A small model was run to enable a Face Evolver to be more quickly developed and evaluated. The precise number of exemplars required for a PCA model to generate faces of acceptable likeness is not known, though guidance may be sought from similar studies. For example, Brunelli & Mich (1995) used 87 faces, Sirovich & Kirby (1987) used 115 faces, Kirby & Sirovich (1990) used 100 faces and Blanz & Vetter (1999) used 200 faces, of which 100 were male. Ultimately, a corpus containing 72 faces were assembled, being much nearer to the number used in these studies, and PCA models were build for shape and texture from them. There is some evidence that this increase resulted in a significantly more complex face space⁵⁸. Note that the first 35 coefficients from the shape and texture models will be used for face generation (i.e. the initial 50% of eigenvectors as before).

⁵⁸ A small pilot study was run with this sized model that evolved randomly generated targets for 4 generations by 15 visitors to the Hatton Gallery, Newcastle. Rating scores (using the AFSS) were collected on the best faces as normal. Other parameters and procedures were the same as Experiment 7 (using 32 population faces); the elitism mechanism was still faulty at this stage. Although the fourth generation increased by an average AFSS rating of 1.0 points, this only approached significance using a *one*-tailed within-subjects t-test ($t=1.53$, $DF=14$, $p=0.074$). It was found however that the average rating scores from this study were significantly lower than in Experiment 7 ($t=3.60$, $DF=154$, $p=0.004$) using a two-tailed between-subjects t-test. Taken together, these results suggest that the face space had become significantly more complex.

The Problem of Unwanted Change in Pose

Although the original images were photographed full-face, small variations in pose are captured by the PCA and become reflected in randomly generated faces. An example of such an effect is shown in Figure 66. Having faces that are created with changes in pose of this nature is likely to be distracting to a witness.

Figure 66: An Example of an Unwanted Pose Change



One solution to this problem is to standardize the corpus images by improving subject pose during photography. This is obviously a time-consuming process and any small errors may still lead to noticeable changes. Another approach would be to “rotate” the corpus faces such that the pose is viewed full-face. It is believed that to perform this accurately, a 3 dimensional model of the head is required. Such a model is outside the scope of the current work and this method is therefore not considered viable.

Although not a full 3D representation, a holistic shape model is of course a component of the photofit system (and arguably a “2D shape” model; as the depth dimension is not explicitly represented). It is suggested that pose correction be performed using this simple shape model. A heuristic for pose re-alignment is based on the observation that the tip of the nose usually lies in the centre of the face and so operations that “re-align” the nose might achieve this rotational effect. One approach is to calculate the horizontal translation necessary to those coordinates that specify the internal features to bring the tip of the nose into alignment. This can be followed by performing a “best fit” operation⁵⁹ of the translated

⁵⁹ In practice, this can be achieved by computing the coefficients in the shape model that have the lowest error for a given set of coordinate points.

coordinates in the shape model. Once the shape coefficients are computed, these can be used to re-constitute a new shape vector. The resulting vector is therefore re-aligned and guaranteed to be within the shape model. Figure 67 illustrates the satisfying effect of applying this process to the face presented in Figure 66 -

Figure 67: Removing the Unwanted Pose Change



The Feature Shifter

The ability to “move” facial features in a specified way is intuitively desirable. It has been observed whilst using the Face Evolver that participants have sometimes wanted to choose a slimmer face or one with the eyes closer together. Indeed, there is evidence that small configural changes made to a face can be very noticeable (e.g. Bruce, Doyle, Dench & Burton, 1991; and Haig, 1984). Haig (1984) has found that a vertical manipulation to the mouth, eyes and nose can be detected even when the movement is close to the visual acuity of the eye. When a desired configuration is not present, the user must rely on a relevant parameter mutation or the selection of an otherwise “poor” face that includes a desired aspect. Either way, this facial aspect may be relatively poorly expressed in the best face, causing possible distress to the witness⁶⁰.

One could of course just perform a “free” morph whereby image distortions are carried out by the uncontrolled movement of pixels: feature manipulations within the *image space*. Such

⁶⁰ Several photofit operators have commented [to me] that when creating a composite, a witness may become apparently “fixated” on a facial feature (such as the hair) to the extent that they cannot focus on any other aspect of the photofit until a satisfactory likeness has been achieved.

operations are of course common in standard electronic photofit systems. Movement in the image space could be problematic in this application though when breeding a new generation of faces: what should one do with the previous pixel translations?

A solution is to follow any feature translations with a best fit in the shape model – as suggested for pose re-alignment. For example, in order to move the eyes together, the horizontal aspect of the pixels that specify the eyes would be brought closer into register and a best fit operation would be carried out in the shape model. A shape vector would then be regenerated (from the eigenshapes) and used to morph the shape-free image to reflect the featural change. As facial features are specified by coordinates, this would allow any features to be manipulated. The solution proposed, perhaps similar to Brunelli & Mich (1995)⁶¹, results in movement within the *holistic face space* [as opposed to movement within the *image space*].

A small utility, known as the *Feature Shifter*, was therefore designed for the EvoFIT system that enabled specified facial features to be moved and resized. An example using this utility that positioned the eyes closer together in the holistic shape space can be seen in Figure 68 –

Figure 68: An Example Illustrating the Effect of Moving the Eyes Closer into Register (by 8 pixels). The original image is on the left



Despite this ability to “navigate” in the holistic shape space, I was concerned whether all young male Caucasian faces could be generated. Part of the concern was that even with a huge database – perhaps containing several thousand images – could an acceptable likeness really be guaranteed? Could a wide face with average ears be specified with a small nose for

⁶¹ Brunelli & Mich (1995) explain that in their “intra feature warping” mode, the PCA expansion coefficients are re-computed given changes in facial configuration, though the mechanism through which this is performed is not specified.

example? An associated issue, is what should be done to faces that have been “damaged” in some way; a broken nose, a black eye and a cauliflower ear⁶² are examples. Whereas the black eye may be resolved as a textural issue, and will be considered later, the broken nose and cauliflower ear are primarily shape distortions. The problem is that without considerable effort in setting up a database containing all possible outcomes of damage, it is unlikely that these representations could be reproduced.

Taken together, these observations suggest that a “free” movement of features should be permitted anyway in addition to movement in the holistic shape space. However, as discussed above, it is difficult to know what to do with these changes when evolving. The best solution appeared to apply any free-morph changes to all faces in the current and future populations. This does have the disadvantage that some changes might be inappropriate to some of the faces, but these faces could be ignored or the free-morph modified appropriately. This free-morph method was therefore also implemented in the Feature Shifter.

Hair and Overlays

In addition to an increase in the complexity of the face model, the other important recommendation made in Chapter 3 was to provide a greater variety of hairstyles. One method would be to take photographs of a large number of hairstyles. This would need to be carried out under controlled lighting conditions so as to limit differences in illumination [between the hair and the face] with light sources originating from different angles. This is an extremely time-consuming process and is arguably unnecessary since such repositories are already available in computerized photofit packages. For example, EFIT and PROfit permit any hairstyle to be saved to disk in one of several common image formats; referred to as “exporting”. This approach is particularly appealing since photofit operators would already be familiar with procedures for hairstyle selection. Allowing operators to make post-selection editing changes in their “favourite” image editor would further capitalize on existing knowledge.

A potential problem with this approach is the correct placement of a hairstyle onto a given set of population faces: not only is the position of the hair important but also its size. The most tractable solution appeared to be to export a reference face [from EvoFIT] into EFIT or PROfit, apply a hairstyle and then re-import it back [into EvoFIT]. If the reference face were to have the average corpus shape, then EvoFIT would be able to treat it as if it were one of the standard hairstyles: keying in the internal features and applying the final morph. In fact, this can easily be achieved if the reference face is one of hairstyles that itself contains little or no hair. All that would be required then would be to take a copy of such a hairstyle, so as not to

⁶² An ear that has been thickened or deformed via repeated blows; typically occurs in boxing and rugby.

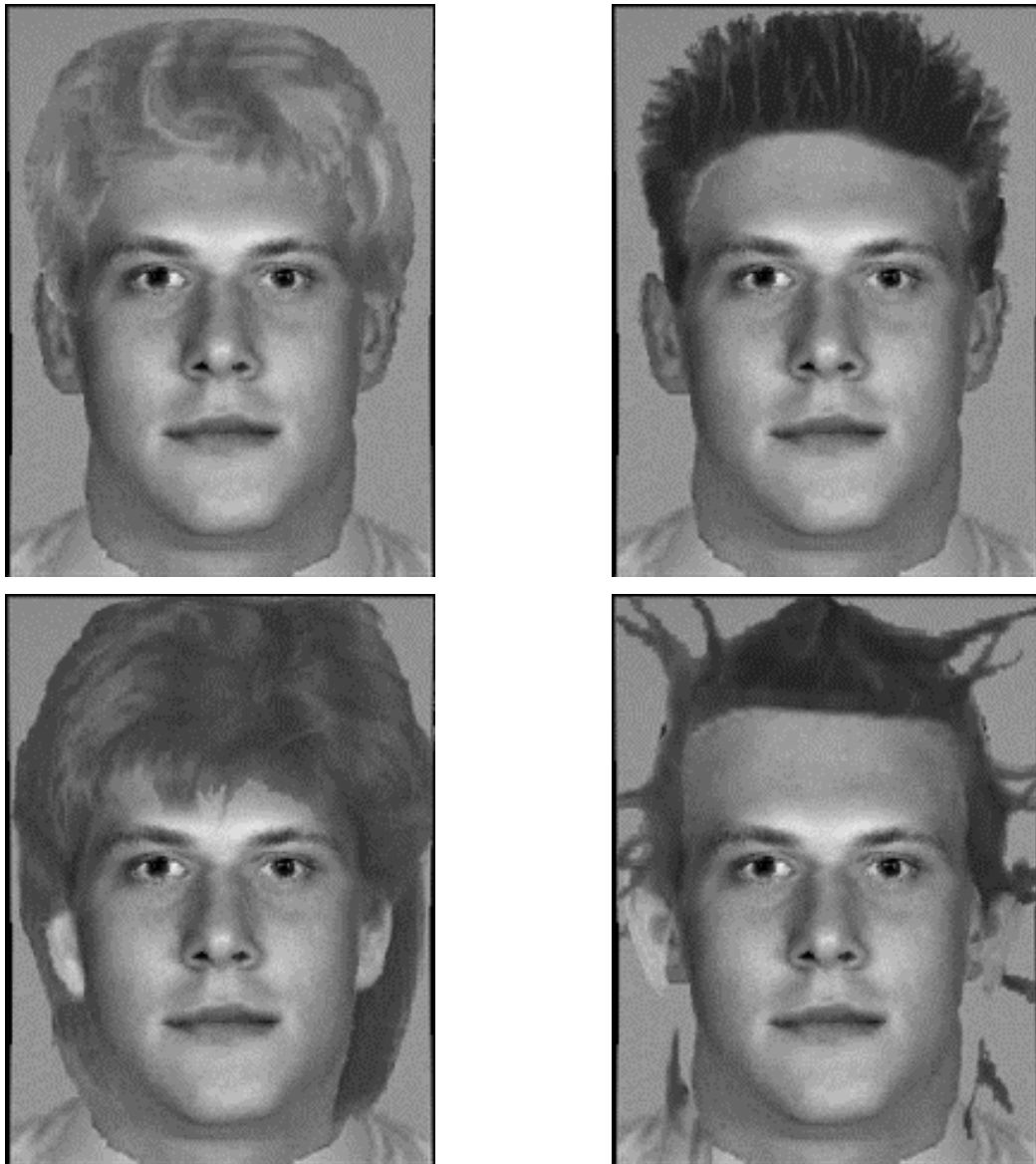
modify the reference images, and to manipulate this temporary or *working* background for the purpose of hairstyle assignment.

Ultimately, due to availability, the PROfit system was used, though provision was made for future use with EFIT. The implemented procedure imported a face from the EvoFIT database containing relatively little hair into PROfit. This was used as a context in which to select, size and position a chosen hairstyle. Other facial features in PROfit – such as eyes, nose and mouth – would be hidden while this is being carried out⁶³. The resulting image would then be saved on disk.

Unfortunately, both EFIT and PROfit apply an image border. A simple utility, called *Import Photofit*, was designed that enabled the manual delineation of the edges to the EvoFIT and ensured that the image size was correct (currently maintained at 180x240 pixels). Some examples of hairstyles extracted from PROfit using this technique are shown in Figure 69. The figure also illustrates the large effect that hairstyle has on facial appearance (refer to Fig. 2.2 in Ellis (1984) for a similar demonstration using Photofit).

⁶³ Instructions for operating PROfit are detailed in Zeda (1998).

Figure 69: Examples of Hairstyles Imported from the PROfit Package and Applied to a Population Face



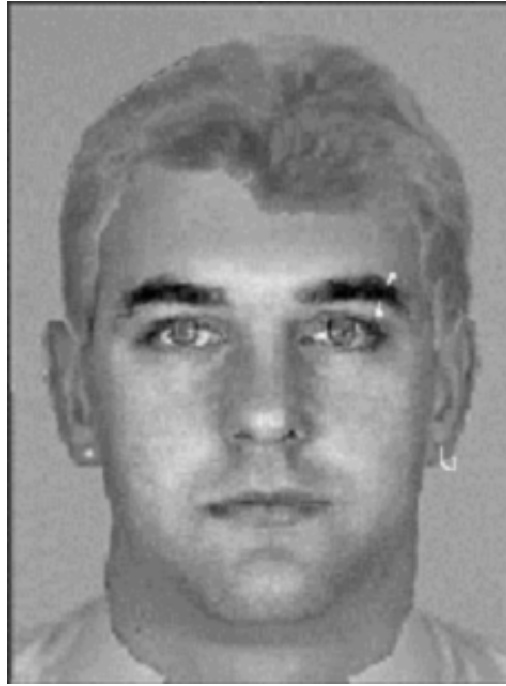
Once the image has been imported, referred to as the *external feature image* (EFI), it could be used as the external features for the current and future population faces. The hair could of course be changed to a different style by repeating the aforementioned procedure.

Provision was made to allow modification of the external features in several standard photographic editing packages: Microsoft Paint, Microsoft PhotoEditor and Adobe Photoshop. The use of these image editors would allow adornments, such as earrings, ear studs and necklaces to be added to the external features, expanding the utility of the EvoFIT system.

A further utility would be to add moles, scars, beauty marks and even adornments to the internal features. The ability to add facial marks is important due the high information content conveyed by them and the associated identification benefit (Zavala, 1972). Once again,

this could be done via an image editor. Of course, editing the EFI would not be a valid solution in this case since the internal features are derived from the texture model. A simple solution is to export a copy of the average shape face, called the internal feature image (IFI), to the editing package. Any editing changes performed on the IFI would be applied to the population faces (by simple numeric addition). An example of adornments added to both the internal and external features can be seen in Figure 70 using this procedure.

Figure 70: An Example of Adornments Added to the Internal and External Facial Features



The functionality necessary to implement these changes to the internal and external features (including hairstyle selection) was combined into a single utility called *Modify Hair & External Features*.

Summary of Developments to the Mark II Face Evolver

In this chapter, considerable development has taken place on the Mark II Face Evolver described in Chapter 3. Development [proposed in Chapter 3] included a larger face model (for both shape and texture information), access to a larger repository of hairstyles (via PROfit) and the ability to edit the external features in a photographic paint package (e.g. Adobe Photoshop). Provision was also made to manually edit the internal facial features [using Photoshop] enabling scars, adornments and other facial characteristics to be added.

The parameter settings and coefficient pruning suggested in Chapter 4 were implemented, along with 2 face palettes that permitted the separate selection of facial shape (FSP) and textures (FTP). Development aimed at encouraging faster convergence continued

with an additional 2 palettes that displayed the *shape* palette (FSP) with a previously preferable texture (FSPBT), and the *texture* palette (FTP) with a previously preferable shape (FTPBS).

Two further software utilities were implemented. Both of these were also aimed at increasing system convergence. The Feature Shifter enabled the relationship between features to be manipulated either within the holistic face space or as an unconstrained relational change. The results were to provide a mode operation not unlike the existing electronic photofit systems. The other utility, the Facial Composite Tool, enabled a new face to be created from the shape of one face and the texture of another.

Algorithms may be found in Appendix H summarizing the generation of population faces and operation of this full system.

Experiment 10: Evaluating the EvoFIT System

The objective of this evaluation was to gain an indication of likely performance were the EvoFIT system to be used by the police at this time. It was considered important therefore that photofits be constructed from a large range of targets and that these constructions be carried out from memory. Of further importance was performance compared with one or more of the current photofit systems. These objectives were satisfied by a collaborative study with Derek Carson from the University of Abertay. This collaboration would involve recognizing a large number of photofits created from memory with Derek employing EFIT and myself, EvoFIT. The study detailed here is the result of the combined design and procedure set out by Derek and myself.

Method

An important aspect of this evaluation is the construction of photofits from memory. The most realistic method would be to allow subjects to view a staged crime taking place and then to create a photofit. In this case, it would be necessary for the “assailant” to be unknown to the subjects. To maintain an ecologically valid evaluation, the resulting photofits would need to be shown to people with whom they are personally familiar. In addition, to obtain a measure of general performance, a significant number of targets should be employed. For example, Kovera, Penrod, Pappas & Thill (1997) created photofits of between 50 and 100 different targets⁶⁴. The problem with staging crimes containing this number of targets is that it is rather time-consuming. As a compromise, it was decided to create photofits of generally well-known

⁶⁴ The methodology of the paper is ambiguous. Two pupils from 5 different High Schools each made 10 composites. But, it is not clear from the paper as to the degree of overlap of targets used between students that came from the same school. This means that somewhere between 50 and 100 of the composites were different.

people, such as famous actors and musicians. Although the use of such stimuli is not *entirely* ecologically valid, given that the targets are likely to be familiar, they can nevertheless be created from memory. The effect of familiarity may not be an issue anyway when constructing photofits from memory, as Davies, van der Willik & Morrison (2000) discovered, though this issue should probably be the focus of a study later on.

Another advantage of using famous faces is that it fits in with current research: Brace, Pike & Kemp (2000) and Davies & Oldman (1999). Both studies constructed composites using EFIT from memory and with the target present. Whereas Brace et al. yielded a recognition rate of 25%, by presenting composites for recognition from both construction modes at once, Davies & Oldman (1999) found that individual composites were recognized only 5.6% in the memory condition and 9.7% when the target was in-view during construction. Using these studies as a guide, one would expect a recognition rate between 5.6% and 25%, though the upper limit is likely to be somewhat lower since Brace et al. presented multiple composites for recognition, a format known to elevate performance (Bennett, 2000; and Bruce, Ness, Hancock, Newman & Rarity, submitted).

It was decided though that subjects be given a short exposure to the target prior to creation of the photofit. This was thought prudent due to reported differences found in recognition studies when varying the retention interval between study and recognition (e.g. Shapiro & Penrod, 1986). All subjects would therefore begin the photofit process with an equivalent exposure to the target and be less dependent on the last time they saw the famous face.

Even controlling for the duration of exposure and the retention interval, it was believed that the overall level of familiarity and distinctiveness of the targets could affect both the creation and recognition of the resulting photofits. Both of these factors are known to affect identification. For example, increases in familiarity have been shown to result in a significant increase in sensitivity towards the internal features (Ellis, Shepherd & Davies, 1979; Young, Hay, McWeeny, Flude & Ellis, 1985; and Davies, van der Willik & Morrison, 2000); and distinctive faces have been shown to enjoy an increase in identification rate (e.g. Shapiro & Penrod, 1986). It was decided therefore to keep the level of familiarity as constant as possible, but to manipulate the level of target distinctiveness.

Selection of Target Stimuli

As mentioned above, one of the objectives of the study was to test performance with a relatively large number of targets. Ultimately, thirty was believed sufficient to gain a good measure of average performance. To ensure that the operators could not initially bias the

construction of the composites, the selection of suitable target photographs was carried out blind to the operators⁶⁵.

These stimuli were obtained first by collecting ratings of familiarity and distinctiveness from 60 written names of famous people. A range between 1 and 10 was used for both scales. For distinctiveness, subjects were asked, "Imagine you are standing on a busy bus station platform, how easily could you pick this person's face from the crowd. Does this person have a distinctive face or would they appear as just one of the crowd?" For familiarity, subjects were asked to rate how familiar they were with each famous person: 1 being unfamiliar and 10 being very familiar. Data from 10 subjects were used. From these ratings, thirty names were extracted and divided into three different distinctiveness levels (low, medium and high) with equivalent familiarity⁶⁶. Reliability was verified by two repeated-measures ANOVAs, resulting in a significant main effect of distinctiveness ($F=107.76$, $DF=(2,27)$, $p<0.001$) but not familiarity ($F=2.26$, $DF=(2,27)$, $p=0.124$). Appropriately, Fisher LSD tests revealed that all contrasts were significant for distinctiveness ($p<0.001$).

Subsequently, good quality, full-face monochrome photographs were obtained (refer to Appendix B).

Creating the Computer-Generated Composites

A procedure used in the UK to train operators to elicit information from a witness is based on a "cognitive approach" (FIC, 1999). This approach, used during a *Cognitive Interview* (CI), is designed to facilitate the recall of as much unbiased information as possible regarding a crime largely through re-instating the context in which the event took place. Part of the CI involves eliciting a verbal description of the suspect, including the face. The verbal description typically involves a phase whereby a witness recalls (and then re-recalls) details of the event in his or her own time with the minimum of external cueing; referred to as "free-recall". This is followed by a more interactive session whereby details about specific events are requested; a "cued recall" (e.g. "What can you tell me about the mouth?"). Following this, a composite would be created and an estimation [by the witness] of the composite's likeness to the assailant would be recorded (a percentage).

To achieve a degree of similarity with real life situations, it was thought best as far as possible to parallel this procedure. Thus, after the exposure of a target, a CI based approach would be used to elicit a description of the face. This would involve two sessions of free recall followed by one session of cued recall. To maintain parallels further, the identity of the targets

⁶⁵ This was performed by one of Derek's Research Assistants at the University of Abertay.

would be hidden from the two operators [Derek and myself], since it is normal for operators not to have prior exposure of an assailant - a procedure adopted previously in research (e.g. Davies, Milne & Shepherd, 1983). In addition to these targets not being known to the operators, subjects would be requested to try not to reveal the identity of the famous person. Whether the identity was revealed either from a subject's comment or from the quality of the photofit, it was though best that the session should continue. Once again, this would serve to parallel "real" photofit situations where no mention is made that an operator might have an idea of the suspect being created. In any case, the role of the operator was understood to control the software tools under the *guidance* of the "describer". According to ACPO(S)⁶⁷ guidelines, "a composite is a pictorial record of a witness's memory and not that of the police artist or facial imaging operator" (ACPO(S), 2000, page 11).

Participants

EvoFITs

Thirteen males and 17 females each created an EvoFIT. Their ages ranged from 15 to 55 and their mean age was 28.1 (SD 9.3). They were paid £10.

EFITs

Eighteen males and 12 females each created an EFIT. Their ages ranged from 18 to 51 and their mean age was 28.9 (SD 9.5). Participation was voluntary.

Apparatus

EvoFITs

EvoFIT software version 2.02e running on a Pentium PII PC clocked at 350MHz was used to create the composites in addition to Adobe Photoshop (version 5.0) and PROfit (version 1.30W). The EvoFIT Faces were displayed on an Iiyama 17" monitor and 30 famous faces were used as stimuli (refer to Appendix B). An EFIT Description sheet was used to record the verbal description (an example of which may be found in Appendix C).

⁶⁶ The mean distinctiveness rating was 6.83 for low, 7.24 for medium and 7.78 for the high distinctiveness condition. The mean familiarity rating was 7.51 for low, 7.85 for medium and 7.94 for the high distinctiveness condition.

⁶⁷ An acronym for the Association of Chief Police Officers (Scotland)

EFITs

EFIT for Windows version 3.1 was used, running on a Celeron Laptop with a 14" monitor. Like the EvoFIT procedure, an EFIT description sheet and a duplicate set of target stimuli were used.

Procedure

The basic procedure was kept the same for the creation of EFITs and EvoFITs. Subjects were told that they would be creating a photofit of a famous face from memory. An envelope was given containing the targets and subjects were instructed to remove one at random. If the person depicted was not familiar, they were told to replace the photograph and select another. When a familiar face was found, 1 minute was permitted for a detailed inspection of the target. The subjects were asked to try not to reveal the identity of the famous person at anytime during the session. The code (on the back of the photograph) was recorded and the target face placed in a second envelope that contained the "used" stimuli.

A short description of the photofit system was provided and an opportunity given for questions. Afterwards, a verbal description of the famous face was elicited, comprising of two cycles of free-recall followed by cued recall; details were noted on an EFIT Description sheet (Appendix C). A photofit was then created using either EFIT or EvoFIT and a percentage likeness was estimated by the subject. The resulting EFITs and EvoFITs may be found in Appendix D.

Recognizing the Composites

Evaluation primarily involved identification rates. This could easily be achieved by asking another set of subjects to recognize the celebrity photofits. However, despite care taken to control for familiarity, concern was expressed that some celebrities may not be well-known (e.g. Michael Owen). Therefore, participants were also asked to name the original targets after they had finished the photofit naming exercise.

As the important aspect was considered to be *recognition* rather than naming ability, an unambiguous semantic description was believed acceptable. For example, "Big lips, oldie, lead singer in 60's band," would be taken as a correct response for Mick Jagger. This approach has been adopted elsewhere (e.g. Lander, unpublished).

The order of presentation was randomized for each subject.

Participants

Eighteen subjects participated, comprising of 4 males and 14 females. These were drawn from students attending the Open University summer school course D209, Stirling University, Stirling (June 2000). Participation was voluntary.

Results

The photofits were recognized a total of 39 times. As there were 540 presentations of the stimuli (18 subjects * 30 photofits), this resulted in a raw hit rate of 7.2%. If one divides the number of times a photofit was recognized by the number of times the corresponding target photograph was recognized, a *conditional hit rate* (CHR) for each photofit may be obtained. This procedure was adopted to further compensate for differences in target familiarity.

The CHR has been calculated and is shown in Figure 71. It can be seen that 13 photofits were recognized in total (43%) and that the CHR ranged from 0 to over 50% (53%); the best recognition occurred for photofits of Nicholas Lyndhurst and Mick Jagger. There were 5 photofits recognized in the low distinctive category, 2 in the medium distinctive category and 6 in the high distinctive category. The average CHR was 9.6% and the average CHR of photofits that were recognized by at least one person was 22.1%.

Figure 71: Conditional Hit Rate (CHR) of EvoFITs Grouped by Distinctiveness (names are sorted by surname within each category)

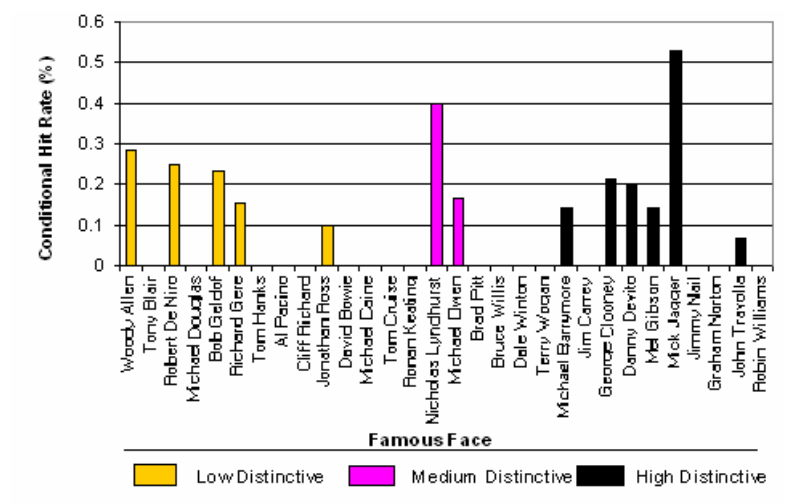
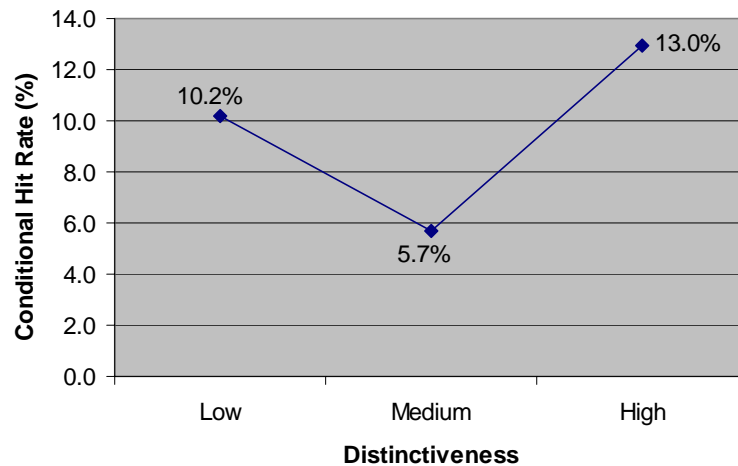


Figure 72 shows the conditional hit rate divided into the 3 distinctiveness categories. It can clearly be seen that the medium distinctness photofits performed worse (5.7%) and the

high distinctive photofits were recognized the best overall (13.0%). The inferential statistics for these data will be conducted in comparison with the EFIT system later in this chapter.

Figure 72: Conditional Hits Rate for Low, Medium and High Distinctive EvoFITs



There was a low, non-significant correlation between the conditional hit rate and the percentage likeness recorded at the end of the photofit session ($r=0.26$; $F=2.07$, $DF=29$, $p=0.161$).

Qualitatively, the EvoFIT operator found that the *Feature Shifter* and *Facial Composite Tool* generally resulted in a likeness (to the target) that was preferred by the subject. In addition, on the second generation, the operator displayed the FSP first, followed by the FSPBT. Appropriately, subjects reported that it was easier to select faces from the FSPBT than the FSP. A similar test was carried out for the FTP and the FTPBS, with subjects once again preferring the FTPBS.

EFITs

Participants

Eighteen subjects participated, comprising of 8 males and 10 females. These were members of staff and students attending the University of Stirling and the University of Abertay. Participation was voluntary.

Results

The EFITs were recognized a total of 88 times. As there were 540 presentations of the stimuli (18 subjects * 30 photofits), this resulted in a raw hit rate of 16.3%. Looking at the CHR

by targets, Figure 73, it can be seen that 22 photofits were recognized in total (73.3%) and that the CHR ranged from 0 to over 60% (61.1%); the best recognition occurred for Woody Allen. There were 6 photofits recognized in the low distinctive category, 8 in medium distinctive category and 8 in high distinctive condition. The average CHR was 17.1% and the average CHR of photofits that were recognized by at least one person was 22.6%.

Figure 73: Conditional Hit Rate of EFITs Grouped by Distinctiveness (names are sorted by surname within each category)

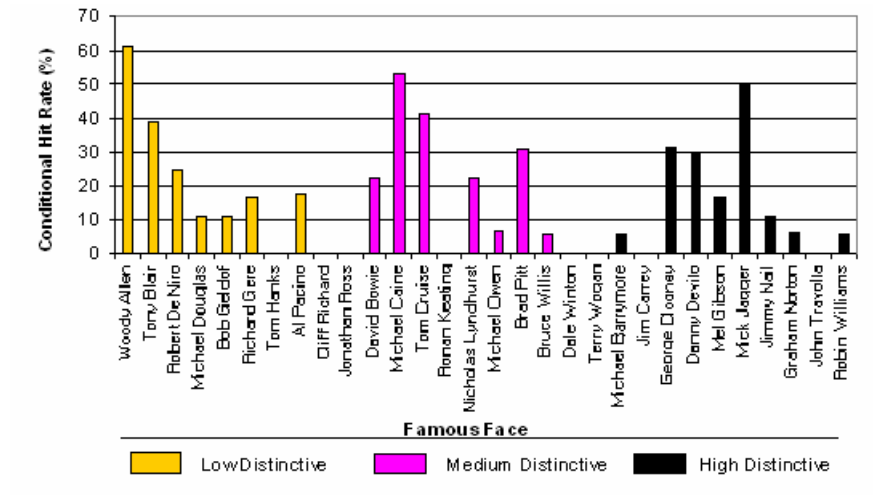
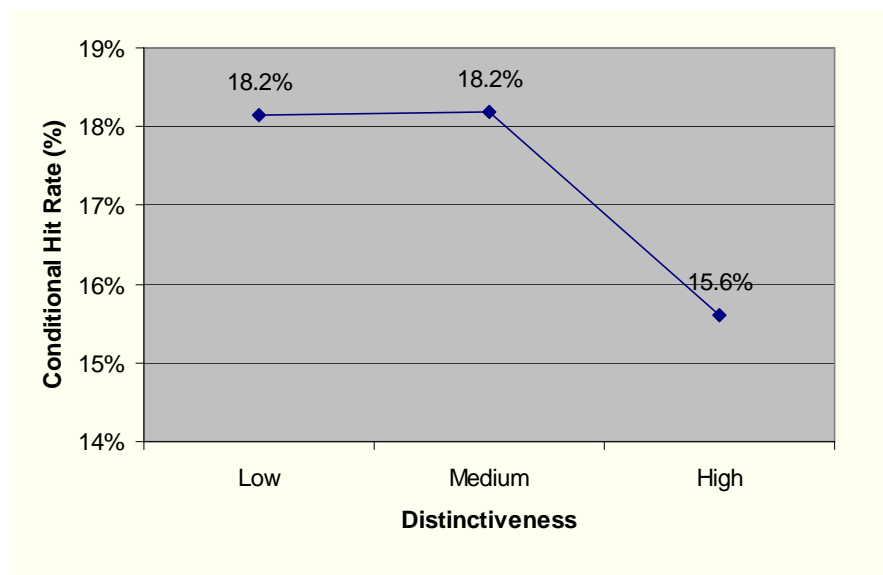


Figure 74 shows the conditional hit rate divided into the 3 distinctiveness categories. It can clearly be seen that the high distinctness photofits performed worse (15.6%) and there was no difference between the other two categories (18.2%). Once again, the inferential statistics for these data will be conducted for both systems later in this chapter (see the following section).

Figure 74: Conditional Hits Rate for Low, Medium and Highly Distinctive EFITs



As before, there was a low, non-significant correlation between the CHR and the percentage likeness recorded at the end of the photofit session ($r=0.14$; $F=0.55$, $DF=29$, $p=0.465$).

Comparison with EvoFIT

A repeated-measures ANOVA indicated that the average CHR scores for the EFITs were significantly higher than the EvoFITs ($F=6.11$, $DF=(1,27)$, $p=0.020$), there was no significant effect of distinctiveness ($F=0.84$, $DF=(2,27)$, $p=0.912$) and no interaction ($F=0.44$, $DF=(2,27)$, $p=0.444$). There is however no significant difference [between systems] in the average CHR of targets that were recognized by one or more people (EFITs are higher by 0.2%; $t=0.04$, $DF=34$, $p=0.970$).

Discussion

In summary, 9 more of the targets were recognized with EFIT and the average CHR was 7.4% higher with EFIT, thus indicating an overall advantage for EFIT. However, 7 out of the 13 EvoFITs (54%) did receive a higher CHR than the corresponding EFITs and there was no significant difference in average CHR of successfully identified photofits. Also, neither system exhibited a significant distinctiveness effect.

It was thought that the last set of changes made to EvoFIT (e.g. Feature Shifter, Facial Composite Tool, and the newer FTPBS and FSPBT palettes) would result in relatively good performance - at least equivalent to EFIT. In reality, the EvoFIT recognition was poor and worse than EFIT. It could be argued that these differences may be the result of differences between operators. It has been found that experienced operators perform better than novice operators (e.g. Davies, Milne, & Shepherd, 1983). However, both operators in the current study were experienced, with at least 10 composites constructed previously. Therefore, although differences may exist between operators, these effects are considered too small to produce the large differences observed.

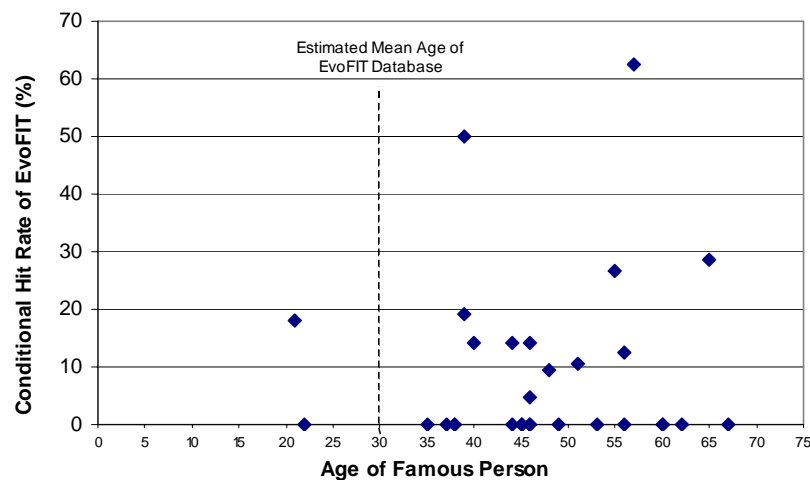
The large difference in hit rates (7.4%) between the two systems is believed to be caused by target age. Several subjects commented during the construction phase that the age of the EvoFIT appeared younger than that of the target. For example, two subjects believed their photofits of Michael Douglas and Robin Williams appeared 10 years younger than in real life, and another believed that their attempt at Cliff Richard needed aging by up to 20 years. A small Internet-based⁶⁸ study with 70 subjects indicated that the average mean age of the

⁶⁸ In this research, 10 EvoFITs (out of a possible thirty) were selected at random for each subject and displayed on a web page. A text box permitted a single age estimate to be made for each image. Data was analyzed from those participants who rated all ten photofits as well as providing demographic information about themselves including their age, gender and occupation.

photofits was 31.6 years (SD 8.8). This was found to be significantly less than the mean age of the targets in this study (mean 47.0 years, SD 11.1; $t=1e6$, $DF=798$, $p<0.001$).

Intuitively, it was thought that the EvoFIT system should be able to construct photofits up to the mean age of the original faces used to construct the face model. As this information was not available directly⁶⁹, a similar Internet-based study with 33 subjects revealed that the mean estimated age was 30.0 (SD 7.9). Interestingly, the mean of the photofits was significantly greater than the mean of the originals ($t=2.93$, $DF=101$, $p=0.004$), indicating that the EvoFIT system is capable of producing photofits beyond the average age of the corpus. As can be seen from Figure 75, however, there were only two celebrities with an age less than 30 years. Of these, only the EvoFIT of Michael Owen was recognized (the other being Ronan Keating). Due to the small number of targets, this study does not therefore represent a measure of likely system performance if used to create a photofit of most suspects, who tend to be in their late teens and early twenties (Goffredson & Polakowski, 1995). The study does nevertheless indicate the likely performance (9.6% CHR) when used to create photofits of targets with an average age well beyond that of the database.

Figure 75: CHR of EvoFITs Ordered by Age of Famous Person



⁶⁹ Age and other demographic data was not supplied with the corpus images obtained from the Home Office.

Experiment 11: EvoFITs of Young Famous Faces

Opportunities became available via a small group project⁷⁰ to investigate whether the EvoFIT system was *capable* of producing age-appropriate photofits. The details presented below are the result of this work. To test the *potential* of the system, EvoFITs were created of young male Caucasian faces with the target-present. It was appreciated that the target-present condition would tend to raise naming rates above the level obtained if creations had been made from memory.

Creation of the EvoFITs

As it was not clear from Experiment 10 whether distinctiveness could be conveyed when the target age was selected more appropriately, EvoFITs would again be created with varying distinctiveness. This was achieved by showing 20 monochrome photographs of young famous faces to 28 subjects and collecting, for each photograph, a distinctiveness rating (from 1 to 10). Subjects were told that the distinctiveness ratings should be based on the degree of unusualness and not on factors such as attractiveness or familiarity. The 5 faces with the lowest average rating and the 5 faces with the highest average rating were selected; the average distinctiveness rating was significantly different between conditions ($t=7.71$, $DF=26$, $p<0.001$). All these faces were recognized at least 75% of the time. The targets assigned to the low distinctiveness condition were Craig Phillips, David Beckham, Noel Gallagher, Leonardo DiCaprio and Matt Damon; the targets assigned to the high distinctiveness condition were Robbie Williams, Michael Owen, David Schwimmer, Stephen Gately and Tim Henman. The mean target age was 27.2 years (SD 4.1).

It became apparent to the “new operators” that the creation of an EvoFIT was a rather complicated procedure. To assist, a set of operating procedures were drafted (refer to

⁷⁰ Comprising of 5 students attending 46AC Cognition, Department of Psychology, University of Stirling, Stirling (Autumn, 2000).

Appendix F) and group members opted to work in pairs to create the photofits. Up to an hour was allowed for each target face. The same system parameters as Experiment 10 were used.

Evaluation of the EvoFITs

The EvoFITs were printed on a separate A4 sheet using a high quality printer. The set was shown to 22 subjects (who had not taken part in the distinctiveness rating exercise and were not aware of the targets used in the study so far).

Results

All EvoFITs were recognized by at least one person and the conditional hit rate was 25.3%. The average CHR for the high distinctiveness group was 33.2% and this was significantly greater than the average of the low distinctiveness condition of 17.4% ($t=3.72$, $DF=21$, $p=0.001$). Following the recognition phase, each subject was asked to rate the similarity of the photofit using the AFSS against the original target photograph. It was found that the average rating for the high distinctiveness group (mean 4.7) was significantly higher than the mean for the low distinctiveness group (mean 3.8; $t=3.39$, $DF=21$, $p<0.001$); the overall mean was 4.3.

Discussion

This small study indicates that the EvoFIT system is capable of producing recognizable photofits when the average age of the targets (i.e. 27.2 years) is more appropriate given the current database. Interestingly, all of the photofits created were recognized by at least one person. Both the recognition rate and the similarity rating scores illustrate a clear advantage when photofitting distinctive faces using this system. Note also that the overall average AFSS rating score for similarity was really quite low (4.3), fitting into the category of "Some similarities", and once again brings into question the use of ratings for photofit system evaluation - especially considering that all the EvoFITs were identified [with relatively few subjects (22)].

Operation Mallard

An opportunity became available towards the end of this research project to undertake a field test of the EvoFIT system as part of "Operation Mallard". This case involves a series of sexual offences carried out by a Caucasian male believed to be in his late twenties in Southern England over the last 2 years (all have been linked by DNA evidence). Sadly, despite considerable effort (including a public appeal), two artists' sketches and a PROfit failed to result in a conviction. Arguably, one problem concerned the likeness of the hair in the sketch

for the third victim (shown in Figure 76). This was due to the victim being unable to mentally form a clear image of the hair. Interestingly, a year later, after having seen a similar hairstyle on TV and then in the street, a clearer image could be formed. An updated sketch was then created for the hair alone and is now believed to be considerably better than the original. It was decided that this updated sketch (Figure 77) be used as a basis for constructing an EvoFIT. Consequently, this sketch was resized, cropped and imported into EvoFIT. The image was then normalized to the average facial shape and the average texture applied to the internal facial features (Figure 78).

Figure 76: Original Artist's Sketch



Figure 77: Updated Sketch of Hair

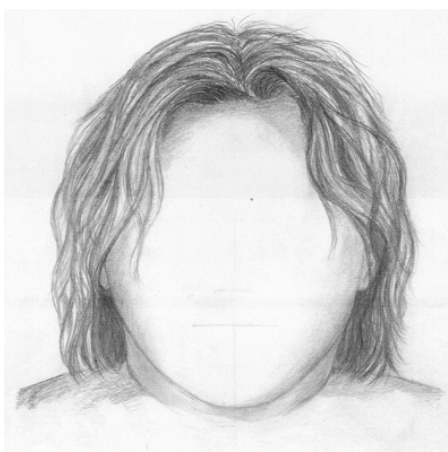


Figure 78: The Updated Sketch Used as the External Facial Features



The construction of an EvoFIT was carried out in the normal way by the selection of shapes and textures using the imported hairstyle. Three generations of faces were required to produce a likeness that was rated at 10/10. The resulting EvoFIT can be seen in Figure 79. The witness has subsequently looked at the EvoFIT and is very pleased with the result. It has been relayed to me that the image has such a powerful effect that the victim finds it difficult to look at.

Figure 79: The EvoFIT Constructed in the Field Test



During the photofit session, the Facial Composite Tool – combining a shape from one face and a texture from another on the Normal Face Palette – was used twice. On both

occasions, a preferable likeness was achieved. The first time, the resulting 'composite' was selected as the best face for that generation. In addition, the Feature Shifter was used once: to reduce the inter-ocular distance by 4 pixels and to close the eyes by 2 pixels. Interestingly, if the EvoFIT is proportionally resized so that the face width matches the width of the original sketch, the shape and spacing between the eyes is almost a perfect match (Figure 80a). If the EvoFIT is then repositioned vertically to align the mouth, the lips and the shape of the jaw is also a very good match (Figure 80b). Notice too that the distance between the mouth and the jaw is also a good fit. In contrast, the length of the face in the EvoFIT is a little longer (10%), as is the nose (10%) and also the distance between the nose and mouth (15%). The nose is a little wider (35%) in the EvoFIT. The eyebrows are shorter and bushier in the EvoFIT and the left eyebrow (as we see it) is positioned slightly higher. Overall, the match is rather good. One does await a conviction, however, to explore the actual degree of similarity between the assailant and the EvoFIT⁷¹.

⁷¹ Analysis is also intended to include not just the similarities between the EvoFIT and the assailant, but also with the other faces selected by the victim during the photofit session.

Figure 80: Comparison of the Spatial Relationship between the Original Sketch and the EvoFIT: (a) Alignment with the Eyes (b) Alignment with the Mouth. The outline of the EvoFIT features is superimposed on the artist's sketch.

(a) Alignment with the Eyes



(b) Alignment with the Mouth



Verifying Anonymity

Although a restricted area of the face space (i.e. the first 35 eigenvectors) was used to prevent the system from generating an exact replica of database images, checks were made to ensure that the EvoFIT was sufficiently dissimilar (see Chapter 6 for a discussion on this point). The first, and arguably the most influential, difference between the EvoFIT and the database images is the hair. All the database images have short hair (coming down at most as far as the top of the ears). None have the same shoulder length hair as represented in the EvoFIT.

The other approach was to compute the error between the EvoFIT and each database image, first for the shape vectors and then for the texture information. For the texture

computation, just the internal features in their shape-free representation was considered. The Root-Mean-Square (RMS) error was used as it provides an average difference from the EvoFIT in *pixels* (rather than pixel-squared with MSE). These differences are plotted below in Figure 81 for shape and Figure 82 for texture.

For shape, the mean error was 6.8 pixels and the minimum error was 3.7 pixels (Face 37). For texture, the mean error was 14.4 pixels and the minimum error was 8.8 pixels (Face 12). Hence, there are significant average internal feature intensity changes (texture) with significant average changes in head shape and facial configuration (shape) between the EvoFIT and the database images. Anonymity has therefore been maintained.

Figure 81: RMS Error for Shape between the EvoFIT and Database Images

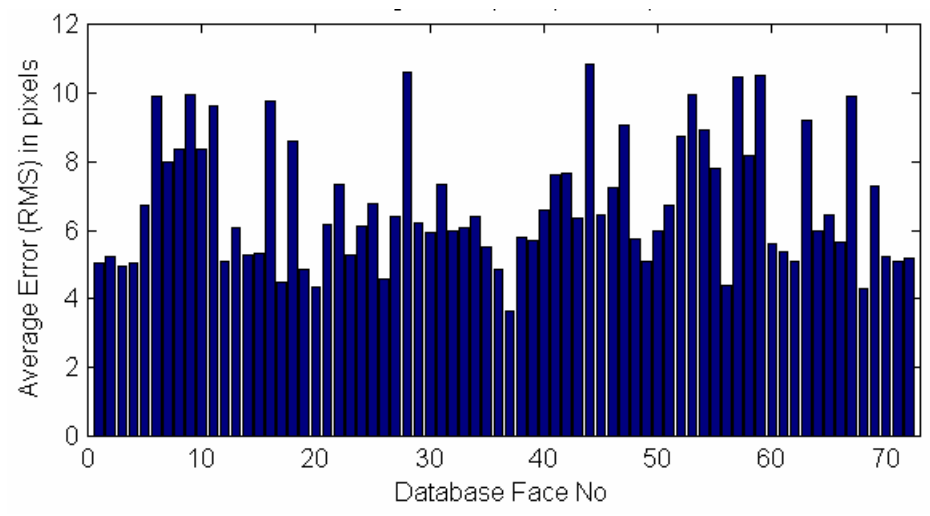
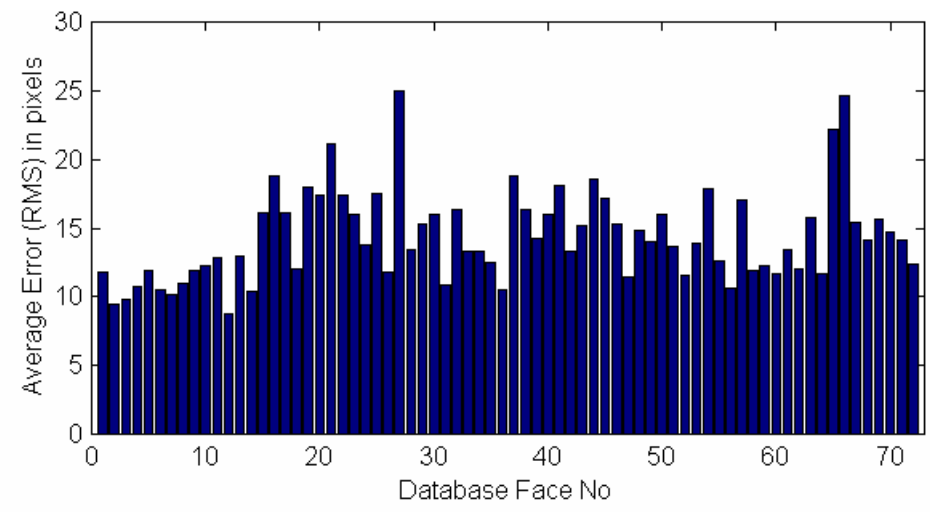


Figure 82: RMS Error for Shape-Free Texture of the Internal Features between the EvoFIT and Database Images



General Discussion

To summarize, the separation of shape from texture selection resulted in a good degree of satisfaction with the resulting photofits. As well as suggesting that photofits created from memory can yield the same rating scores as those created with continual reference to the target, Experiment 9 also highlighted the huge effect of the hair on rating scores. With further development, a full photofit system was created. This allowed the selection and modification of a wide range of hairstyles. The effects of noticeable changes to head pose were removed. Tools were added that enabled facial features to be “moved” both within the image space and also the holistic face space. Two more face palettes were added, attempting to provide a more “convergence-based” representation (adding best texture to the FSP and best shape to the FTP).

During the creation of the 30 famous faces in Experiment 10 and during the Field Test, subjective feedback from subjects was found to be positive regarding the addition of these new palettes and software tools (the Feature Shifter and the Facial Composite Tool). Of course, the precise effect of these enhancements does await further, more quantitative analysis. This could be carried out, ideally using recognition rate as a dependent measure, by the creation of photofits with and without the specified software tool or palette. Note that there are important issues surrounding the Feature Shifter and these are discussed in the following chapter.

Comparing systems, celebrity EFITs were better recognized than the EvoFITs (17.1% vs 9.6% CHR). Interestingly, the EFIT performance turns out to be very similar to that found by Davies et al. (2000). This indicates that naming rates of about 17% are likely with composites constructed from memory using this system. The data also suggests that EFIT does not exhibit a distinctness effect. Of course, one should investigate whether this result still holds when a witness’s memory component is removed. One sensible approach would be to use the stimuli and basic procedure from Experiment 11 to evaluate EFIT.

It was hypothesized that age was a limiting factor resulting in a relatively poor recognition rate for EvoFIT. To determine whether this was correct, a follow-up study (Experiment 11) created EvoFITs of 10 targets with an average age of 27 years. To reduce witness memory effects, the study constructed EvoFITs with the target present and does not therefore constitute an ecologically valid method of construction. The recognition rate (CHR) was found to be 25.3% and there was a strong effect of distinctiveness, mirroring expectations in human recognition performance. This shows that EvoFIT has the potential to create recognizable composites. The data also compares favourably with Brace, Pike & Kemp (2000) who found a naming rate of 25% with their EFITs of famous faces using a presentation format involving composites also constructed with the target present.

Clearly, Experiment 11 shows that EvoFIT is able to create recognizable composites. Despite EvoFITs being constructed more realistically in Experiment 9, from the memory of an

unfamiliar target, evaluation has not been carried out in this more realistic way for the full system (e.g. with improved hairstyle selection, new palettes and software tools) with composite naming. Before such a trial is conducted, the effectiveness of the EvoFIT system in a more realistic situation is unknown. A logical next step is to build composites from the memory of unfamiliar (famous) targets, with evaluation via spontaneous naming.

Analysis was also carried out on the percentage likeness scores attributed at the end of a photofit session. No significant correlation was found between the ratings and the CHR for either system. In studies that have examined subjects' confidence and their actual ability to recognize a target, mixed results have been found. As Thomson (1995) points out, this relationship "has ranged from negative, nonexistent, spurious, weak, to moderate" (page 140). It would seem that only in the more critical situation of identification parades, has a modest overall correlation between confidence and recognition been found (from a meta-analysis by Sporer, Penrod, Read, & Cutler, 1995).

Conclusion

Although the overall identification rate of the celebrity EvoFITs constructed from memory was low, and less than the corresponding EFITs, target age was found to be a factor. When the targets were selected more appropriately, the ability to better represent distinctive targets indicates that the system is sensitive to facial salience. A subsequent Field Test with an age appropriate target is clearly the most forensically relevant: construction of an unfamiliar face from the memory of a crime. A follow-up study should now be conducted to explore performance of EvoFIT when used in this realistic way.

Chapter 6: Future Work

The development and research to date indicates that further research is appropriate for EvoFIT. This chapter details the areas of research and development considered necessary to further the approach. It is argued that important work should be directed at the evaluation of EvoFITs constructed via more appropriately aged targets from memory and the assurance that these images are sufficiently dissimilar to the faces in the database. Research is also required to confirm database size/composition and the effect (if any) of photofit construction following a verbal description. Proposed developments are also discussed that include a different presentation format for the Feature Shifter utility (a parallel interface) and the ability to more rapidly modify facial textures (such as eye colour). Other main developments include the construction of a ¾ view database and the creation of EvoFITs from multiple witnesses.

Summary of Previous Work

The Face Evolver has completed many cycles of development in this thesis. The initial design (Mark I Face Evolver) modelled the gross shading and pixel intensity changes (*texture*) to a face and found a significant decrease in error scores (MSE) with increasing generation. The configural changes between facial features normally found in human faces were then added along with external features, especially the hair, in the Mark II system. A significant improvement in the quality of the images generated (obtained by user ratings) was also found with increasing generation and population size, though a decrement was observed when targets were evolved that were obtained external to the face model (famous faces). A series of simulations provided a set of more appropriate parameter settings as well as techniques believed to facilitate faster photofit convergence (i.e. the separate selection of shape and texture, and coefficient pruning). Six more facial representations (palettes) were then added to further separate the selection of shape and texture; a tool was devised to “move” facial features on request; and the ability to select large repertoires of hairstyles (and modify them in standard photographic editing packages) enabled the EvoFIT system to be created. The evaluation of this final software version resulted in a spontaneous naming rate of 9.6% for famous faces constructed from memory. This low level of performance was believed to reside in the targets being beyond the age capability of the database. A follow-up study revealed that system performance could be considerably higher if appropriately aged targets are used, though this evaluation was not conducted from the witness’ memory and therefore does not constitute a realistic scenario.

Clearly, significant research is required to evaluate EvoFIT before it can be claimed to perform well in a forensically similar situation and a product is made commercially available. There are three main areas believed necessary: the evaluation of young male composites constructed from memory, the appropriateness of the face model and issues surrounding the

verbal description and Feature Shifter. These are discussed in the following sections together with additional developments likely to improve performance.

Evolution from Memory

The last experiment of this thesis (Experiment 11) demonstrated that the EvoFIT system was capable of creating more recognizable photofits when the target age was more appropriate. Another finding was that EvoFITs could convey facial distinctiveness. This is a nice result as it is known that distinctive faces are better recognized (e.g. Shapiro & Penrod, 1986). Insensitivity to distinctive faces would indicate a deficiency in either the process used to create a photofit or in the system itself. Derek Carson's study finds no evidence that the EFIT system is able to convey facial distinctiveness when composites are created from memory. Sadly, no data is available to determine whether this deficit lies in the EFIT system or in the mode of construction (i.e. composite construction from memory). Nevertheless, a follow-up experiment is considered essential to ascertain the overall level of system performance (and also the effect of distinctiveness) using age appropriate photofits constructed with the EvoFIT system from memory. As data is available with constructions made with the target visible (Experiment 11), providing baseline data, this notion could initially be investigated by repeating this experiment with constructions made from memory (as in Experiment 10).

An associated issue when continuing to evaluate the EvoFIT system refers to the use of *unfamiliar* rather than *familiar* (or famous) faces. This is likely to be an important issue since composites are normally created of individuals who are not known to a witness. There is considerable research demonstrating that familiar and unfamiliar face perception is very different (Bruce, 1988; Ellis, Shepherd & Davies, 1979; Hancock, Bruce & Burton, 2000; and Kemp, Towell & Pike, 1997). The consequence of a more fragile facial representation in the realistic (unfamiliar) mode of construction is a poorer quality composite. One would expect therefore to find a decrement in EvoFIT performance when switching to unfamiliar targets. Davies et al. (2000) did not find such a difference and this may be due to the insensitivity of the EFIT system when constructions are made in a "forensically friendly format" (i.e. constructions carried out from memory).

This evaluation could be achieved by creating EvoFITs in a location where target familiarity can be manipulated. Such an experiment could be carried out with relative ease between universities or university departments (e.g. Davies, van der Willik & Morrison, 2000; Bruce, 1982; and Bruce, 1986).

Face Model Issues

Representation

Research is also considered necessary for the face model itself. Even with the enlarged model, built from 72 faces, it is still unclear whether this model is sufficient in number and/or representation to be able to generate all young, male Caucasian faces with an acceptable likeness. One method to guard against this possibility was the “free morph” mode of the *Feature Shifter*, allowing facial features to be moved independently of the face shape model. However, it is not clear at this time when and where a free morph should be applied.

Arguably, the easiest method to establish a suitable model size and composition is by simulation. This could be achieved by assembling a relatively large number of further target faces. This is perhaps best achieved using the Home Office collection (i.e. the image repository from which the EvoFIT face models were constructed), as lighting and pose have been controlled in the same way as the other database images. A “best fit” analysis on this “test set” could then be carried out with databases of increasing size and membership, indicating if there are there any faces which are relatively more important. Although best fit analyses have been carried out with completed databases (e.g. Blanz & Vetter, 1999; Craw & Cameron, 1991; and Troje & Vetter, 1996), no formal analysis of the type suggested here is known.

An approach based on Troje & Vetter (1996) shows promise however. They have adopted a “leave-one-out” analysis such that each face is systematically removed from the database, a PCA rebuilt and the error computed in reconstructing the “omitted image”. They were able to demonstrate that a “testing error” as low as 6% for shape and 12% for texture with their database size of 100 faces. Such an approach may be valuable in determining database membership.

Colour

An associated issue is database image mode: should colour be used? Of course, a face model was constructed in monochrome primarily due to the large increase in model size that would have resulted if hue information had been included. As mentioned in Chapter 2, colour information is not necessary for face perception (e.g. Davies & Thasen, 2000; and Kemp, Pike, White & Musselman, 1996). Further, research does suggest that in matching tasks carried out from memory, colour stimuli can actually result in an increase in false alarms, while leaving the hit rate unaltered (Davies & Thasen, 2000). The overall effect appears to increase a subject’s confidence in selecting a wrong person from a line-up. While this may seem undesirable generally, it may be of value in the EvoFIT system. If the effect of colour is to increase the chance of similar looking foils being chosen, then the task of selecting population faces may also be increased. This benefit may of course be offset by a decrement in the ability to

ultimately select a population face as the photofit. Note also the overhead of locating a relatively large collection of full-colour hairstyles. Nevertheless, there appears to be sufficient justification in developing and evaluating a colour system.

$\frac{3}{4}$ View

Another development likely to be of value is a $\frac{3}{4}$ view database. As discussed in Chapter 2, this notion is based on significant research indicating a recognition benefit of a $\frac{3}{4}$ view compared against profile and full-face views (e.g. O'Toole, Edelman & Buelthoff, 1998; Patterson & Baddedly, 1977; and Shapiro & Penrod, 1986). One reason for this is that a $\frac{3}{4}$ pose better represents the depth of facial features, especially the nose (but see Bruce, Valentine & Baddeley (1987) for a discussion on this point).

Significant work has already been conducted using PCA models based on 3D shape information rather than shape derived from "flat" photographs (sometimes known as *2D shape*). The result is a full 3D representation of the head (e.g. Blanz & Vetter, 1999; and Troje & Vetter, 1996). However, there are simpler methods though available in the public domain for creating different views. For example, Lanitis & Cootes (1997) provide up to a 45 degree facial rotation from an algorithm trained on a frontal pose. For an even greater rotation, Cootes, Walker & Taylor (2000) provide an algorithm trained on just 5 pose angles. Simply, these algorithms could be used to create a $\frac{3}{4}$ view as a post-processing stage following the generation of the population faces. One problem with this method is how to modify the profile information. For example, what should be done to correct a nose whose aspect is too prominent?

Another approach currently being considered in the Psychology Department is to employ a commercially available software package to fuse frontal and profile views into a desired pose (3DMeNow by BioVirtual, 2001). The texture and shape information would be derived from PCA models of simultaneous front and profile views. This idea has the immediate advantage that the full-frontal information (for shape and texture) would have already been prepared and only profile information need be pre-processed (by the alignment of coordinate points and a shape normalization of the image texture). This notion is more than just a speculative idea as development potentially using the BioVirtual software is due to start this year (2001) as part of the current CRIME-VUs⁷² research project in the Psychology Department, Stirling University.

Caution should be applied when evaluating the $\frac{3}{4}$ view database by the construction of EvoFITs of unfamiliar targets. The use of full-face target photographs may now not be appropriate due to an inherent change in pose; the adverse affect of pose on unfamiliar face processing is well documented (e.g. Bruce, 1982; Davies & Milne, 1982; and Hill & Bruce, 1996).

⁷² An abbreviation for *Combined Recall Images from Multiple Experts and viewpoints (VU)*.

This suggests that $\frac{3}{4}$ view stimuli should be used anyway, although multiple views, live video or – best of all – a staged event may represent more ecologically valid conditions. Certainly the mode of target presentation is an important factor in face recognition (Davies, 1983b; Shapiro & Penrod, 1986; and Shepherd, Ellis & Davies, 1982).

Additional Databases

In addition to composition and mode, it may be considered of value, with considerable effort, to add further databases. It was suggested in Experiment 10 that the face model was a limiting factor in the recognition of the famous faces. This also suggests that the addition of an older database may be of value. One might also consider providing a database for female and non-Caucasian faces. Although databases from traditional systems tend to be separate, it is not clear at this stage whether databases for the EvoFIT system should be mixed gender and/or race. Of course, Troje & Vetter (1996) and Blanz & Vetter (1999) have developed a mixed gender database with apparent success; and Baker & Seltzer (1998) has performed a mugshot search on a large database with varying age, gender and race with considerable success. The question requiring address is what might be gained with mixed composition corpora? One immediate advantage is that a mixed race model would provide representation for persons whose parents were of different ethnic backgrounds. Though, it may be very undesirable in a photofit setting for mixed race faces to be generated when the target background has been established. A similar argument can be made for gender and age. Therefore, unless sufficient control is ensured over image generation, mixed composition databases are undesirable (unless a case specifically requires it).

Anonymity

In contrast to composition, development is also necessary to prevent the original faces in the database from being evolved. The maintenance of anonymity has been observed right back to the early development of the Photofit kit by sampling features from faces rather than the inclusion of all features. In Penry's own words, "From one picture I 'borrowed' the nose, from another the eyes and from others, the mouth, chin or forehead/hair" (Penry, 1974, page 4).

Clearly, it is not possible to adopt this strategy with EvoFIT as the representation is holistic and necessarily requires not just the complete set of internal facial features but information about head shape as well. An equivalent safeguard though would be to prevent evolution to the database's shape and texture coefficients. The problem with this approach is that floating-point values are used for coefficients and therefore one may create an exceptionally good likeness to a database image with slightly different coefficient values. One solution might be to prevent images from being generated from within a fixed distance of the

originals; creating a so-called “hypersphere of protection” around the database images in face space. Alternatively, that the population faces are generated from a truncated face model may itself provide a sufficiently different representation to the original images so as to render further attempts unnecessary. This is based on an observation by Sirovich & Kirby (1987) that truncating the Karhunen-Loeve (LV) series results in a measurable error when constructing the original face data.

Recall that an investigation was carried out when the EvoFIT system underwent the current field test. In addition to large differences in hair, it was also found that the database images exhibited very different shapes and textures to the resultant EvoFIT. Of course, it is preferable to apply all necessary checks during the composition stage, rather than thereafter, and further work is therefore required in this area.

Non-Holistic Bias

Verbal Description

A further avenue of investigation concerns the effect of eliciting a verbal description from a witness. As mentioned previously, the description is required prior to building a composite with the traditional systems. It enables a photofit operator to select an initial set of facial features. Further verbalizations follow that describe what is believed to be wrong with the composite. Clearly, neither of these two descriptive components is fundamental to the EvoFIT system. Indeed, it may be the case that the production of a description reduces the ability of a witness to correctly recognize when a set of optimal features has been found. This notion is borne out of research suggesting that verbalization of a target, non-target or another non-face stimuli results in a significant decrement in face recognition (e.g. Dodson, Johnson & Schooler, 1997; Fallshore & Schooler, 1995; Meissner & Brigham, 2001; Schooler & Engstler-Schooler, 1990; Schooler, Ohlsson & Brooks, 1993; and Westerman & Larsen, 1997); the interference to face perception caused by verbalization has been appropriately termed “verbal overshadowing” (Schooler & Engstler-Schooler, 1990). The research also suggests that the phenomenon is less likely to occur if a delay is inserted prior to a recognition task (Messner & Brigham, 2001). For example, Finger & Pezdek (1999) found alleviation from overshadowing after only 24 minutes delay. . There is also evidence that the conveyance of the verbal description itself results in composites that are less recognizable by others (Brace, Pike & Kemp, 2000).

A reason for an adverse effect on recognition is that verbalization results in a reliance on featural information, such as shape of the eyebrows and chin (e.g. Dodson, Johnson & Schooler, 1997). It may be the case therefore, that the verbalization of a face may be detrimental when used in conjunction with the EvoFIT system. It is the case then that - with the exception

of the hair, gender, age and race - no verbal description is necessary in this system (though, as mentioned earlier, with a mixed gender, age and race database, even these descriptors may be redundant). It would appear highly prudent therefore to compare EvoFITs constructed by participants who have generated a verbal description and by those that have not (and maybe also by inserting a significant delay after verbalization or encouraging a faster selection of faces). A further experiment that is likely to reinforce the holistic nature of the EvoFIT approach would be to manipulate the encoding instructions. Indeed, one would expect preferential results from subjects that attempted a facial coding via a trait attribution (holistic) rather than by a more physical facial examination (feature-based); an effect opposite to that observed with standard photofit systems (e.g. Wells & Hryciw, 1984; and Laughery, Duval & Wogalter, 1986).

Feature Shifter

A related, but important question arises as to the efficacy of the tool that manipulates the relationship between facial features: the *Feature Shifter*. The first concern is whether its use does result in a facial configuration that is perceptually closer to the target. There is some supporting anecdotal evidence for its utility from the field test described in Chapter 5. Recall that the use of this tool resulted in a rather good match for eye shape and inter-ocular spacing (comparing the EvoFIT with the original artist's sketch). However, a more formal analysis could be carried out by extracting those faces in Experiment 10 before and after the tool had been used and asking a different group of subjects to choose which ones were more like the celebrity face. If the Feature Shifter were working, one would expect the manipulated images to be perceptually closer than the unaltered ones.

The other major issue with the Feature Shifter is that it may bias the cognitive system towards non-holistic face processing by engaging in a feature-based activity. There is good evidence to suggest that recognition deteriorates following a non trait-based activity prior to recognition (e.g. Berman & Cutler, 1998; Bower & Karlin, 1974; Dodson, Johnson & Schooler, 1997; and Schooler & Engstler-Schooler, 1990). The consequence is that a witness's ability to subsequently select population faces (and ultimately pick a final photofit) might be likewise adversely affected following the use of the Feature Shifter. One way to test for adverse affects would be to compare the pattern of selection made by subjects who had used the utility and those that had not. It is already known that subjects overlap at least 90% of the time in their selections⁷³ and if the utility were detrimental, this figure would be expected to reduce. Were it

⁷³ In a small pilot study, 15 subjects were shown the same set of 16 randomly generated faces together with a target face. They were instructed to select the six faces that were most similar to the target. It was found that the top 6 faces selected most often accounted for over 90% of the total selections (82/90 = 91.1%).

found to be detrimental to performance, restorative measures could be considered that are known to facilitate recognition, such as the re-instatement of context and visual rehearsal (e.g. Shapiro & Penrod, 1986; and Sporer, 1988).

Given the danger of encouraging feature-based facial processing, it may be better to avoid using the Feature Shifter in the mode prescribed. In fact, its use parallels the process by which composites are created in the main electronic systems (e.g. Mac-a-mug, PROfit, and EFIT). A better interface would be based more on the recognition ability of a witness; achievable of course by a parallel interface – like the window used to display population faces (i.e. the *Face Palette*). It is imagined that such an interface would display a range of possible configural changes to a given feature for selection. For example, if a change to the inter-ocular spacing was required, a number of examples could be displayed simultaneously with varying horizontal spacing of the eyes.

A potential problem with this method is in determining the amount of change to apply between successive examples displayed on this parallel interface. The problem is that if the presented manipulations are too small, a witness may not only find the task irritating but be needlessly exposed to facial information. Now, it is well-established that different facial features, or groups of facial features, have different salience for unfamiliar faces (e.g. Ellis, Shepherd & Davies 1979; Matthews, 1978, Walker-Smith, 1978; and Young, Hay, McWeeny, Flude & Ellis, 1985). Interestingly, Haig (1984) has determined the manipulations necessary to an unfamiliar face before changes become noticeable (i.e. the Just Noticeable Difference or JND) in a wide range of facial features. This should enable calibration of facial feature manipulation to be carried out. Note however that there is evidence that the granularity of feature changes proposed by Haig may not be entirely correct, since a significant shift in performance has been observed with configural changes to the hair and eyes for newly learned faces (Honeyman, unpublished data; and O'Donnell & Bruce, in press).

Anchored Face Similarity Scale

A further associated issue concerning non-holistic bias is the Anchored Face Similarity Scale, or AFSS, used to evaluate the quality of a composite. Although there is reason to believe that the scale is consistent between subjects (Experiment 3), it could be argued that scale categories “few similarities”, “some similarities” and “many similarities” result in a featural bias due to the reference made to numeric quantities (i.e. “few”, “some” and “many”). Its use may consequently, like the Feature Shifter, result in a worse ability to select population faces. To date, there is no anecdotal evidence indicating that EvoFIT subjects were worse as a result of the AFSS (as observed by the EvoFIT operator). Clearly, it is the case that good performance can be achieved in some subjects even though the AFSS is used: the EvoFITs of Nicholas Lyndhurst and Mick Jagger (Experiment 10) are examples where at least 40% recognition can be

achieved. Nevertheless, the scale should either not be used with EvoFIT or formally evaluated. For the latter, one could explore whether an individual's ability to select faces would be adversely affected following the use of the scale. One could, for instance, record selections made by subjects for the same set of population faces (given the same target) before and after AFSS use. If there was a similar *overlap* in selections with and without AFSS rating, it is unlikely that the scale would be adversely affecting performance.

Holistic Theory

Central to this thesis has been the notion that faces are perceived holistically and the ability to select similar looking faces [to a target] is preserved if faces are not segmented into their facial features. A sensible question then is whether the EvoFIT approach provides evidence that we perceive faces holistically? Irrespective of how well EvoFIT might or might not perform, the answer to this question is unclear. It is clear though that the basic process driving EvoFIT, the selection of faces, is a holistic operation: population faces do not require to be explicitly segmented into features for selection. There is also no need to describe *why* a population face is (or is not) preferable [to the target], a process that may bias non-holistic processing. Unfortunately, other activities involved in composite construction may result in a non-holistic bias, as already mentioned above: the production of a verbal description, the use of the Feature Shifter and rating with the AFSS. Clearly, additional research is necessary to establish the effect of these potential confounds before informative comments can be made regarding the holistic nature of EvoFIT.

Another potentially useful avenue of research is the effect of target encoding on EvoFIT performance. Recall that Wells & Hryciw (1984) found that instructions suggesting a feature encoding led to better Identikits than a holistic encoding. This would be a valuable experiment applied to the EvoFIT system. Were EvoFIT to be truly capitalizing on the holistic nature of face processing, one would expect a reverse trend, with better EvoFITs produced following a holistic type of encoding. If this were to be the case, such information might also be a useful in a criminal investigation (especially if performance was markedly better than current composite systems under similar conditions). Were a witness to be demonstrating a holistic bias in their memory of an assailant - perhaps referring to the perpetrator with personality traits - an EvoFIT might be an appropriate composite tool.

Enhancing Performance Still Further

Simulations

Recall that in Chapter 4, simulations were carried out with the randomly generated targets. This was done for convenience. It is assumed that the results are applicable to all faces

evolved with the system. Nevertheless, simulations with externally-derived targets should be run; these could be drawn from the “test set” mentioned earlier. It is likely that system performance would be noticeably worse anyway with faces not in the database - as Troje & Vetter (1996) have found.

A different simulation approach is proposed however. It was noted in Chapter 4 that there is a tendency for parameters in an evolutionary system to *interact* with each other; running a set of simulations that evolve a single parameter may not therefore give the best indication of settings. A much better solution would be to *evolve* the parameters themselves! This can be achieved by breeding solutions together, where each solution is a combination of evolutionary parameters. Evolution would continue until a superior set of parameters was produced. Such an approach has already been found to be of value in Caldwell & Johnston’s (1991) composite approach with a “meta-level GA” that optimized the mutation and cross-over rates.

A further modification to the simulation process would also allow parameters to change their settings during evolution. It has been observed previously that a parameter setting may only be valuable for a limited number of generations (e.g. refer to Figure 46, Figure 56, and Figure 57). Allowing parameters to be “disabled” or otherwise changed might therefore increase the rate of convergence to a target.

Feature Shifter

Originally, the Feature Shifter was planned to manipulate both the relational and featural aspects of a face. Of course, the former was achieved, though a lack of time prevented a similar process to be carried out for the texture model. It was planned that manipulations in the texture face space would enable simple processes like lightening the shade of the eyebrows. In more detail, this would involve increasing the intensity of the pixels that comprise the eyebrows and then performing a best fit in the texture model (rather like moving facial features in the shape model). Up to now, if a user has required not just a lightening, but also a darkening of any feature, this has had to be carried out by modifying an overlay mask loaded in Adobe Photoshop. In similarity with the shape model, this was deemed necessary to account for any faces that were not captured by the statistics of the face model. For example, it is unlikely that a statistical representation would be available for David Bowie (Experiment 10) since the colour of his irises are unusually different from each other.

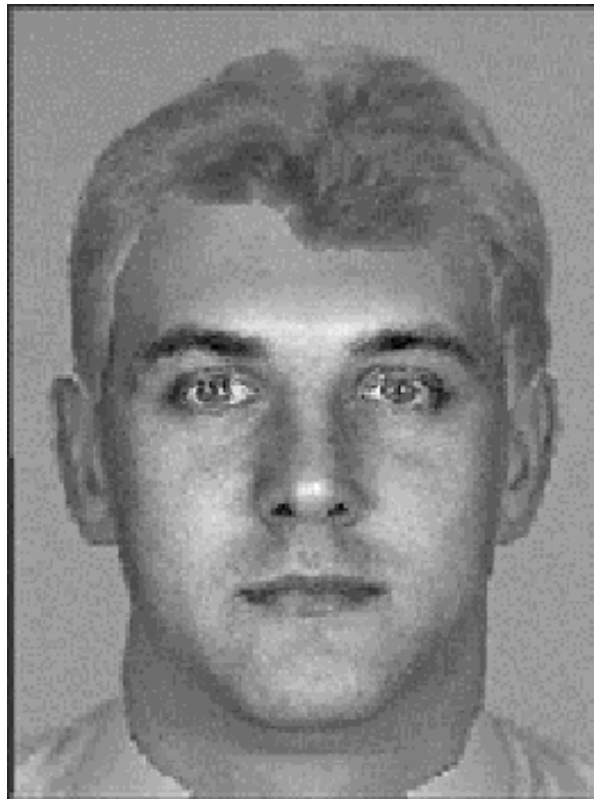
An associated system development might also include the pre-specification of feature intensities. For example, it is frequently the case that people with light coloured hair also have light coloured eyebrows. Therefore a witness may have provided such information as part of a verbal description (or report this information while viewing a population). Being able to automatically bias the population in this way might therefore prove valuable. Effects such as

these could be achieved in a similar way that minor misalignments in pose are corrected: by repeatedly manipulating a feature and perform best fit in the face model until a desired setting is reached.

Eyes and Hair

A very noticeable area of development concerns the eyes. The problem here is that the morphing necessary to create the variations in spatial relations normally seen in faces has the consequence of distorting the iris of the eye. An example can be seen in Figure 83 where the normally round irises have become inappropriately elongated in the horizontal direction. Although to date, this has been overcome by adjustment in Adobe Photoshop, such an approach is in general impractical.

Figure 83: Imperfections in the Generated Images: distortions to the Irises caused by Morphing and the Problem of “Floating” hair



One of the reasons for the distortion lies in the coarse scale of the shape model. When the shape model was first constructed, it was believed that integers would be sufficient to specify coordinates. This assumption appears to be largely correct for all facial features except the eyes; it was found that a single pixel movement was often sufficient to “correct” for a distorted iris. This suggests that a shape model comprising of floating-point values should be constructed. However, this is unlikely to provide a complete solution since not all types of eye

will be readily represented. For example, what should be done when the iris is positioned high in the eye socket? As the picture of Robbie Williams (Figure 31) illustrates, there is sclera now visible between the bottom of the iris and the lower eyelash. Similar problems can be seen with a “lazy” eye or an artificial eye, where the eyes tend not to move together. Effects such as these may be rather difficult to achieve using the current face model.

A preferable solution would be to treat the eyes as an independent feature, rather like the hair. A likely method of implementing such an approach would be to automatically select exemplars from a repository of eyes (e.g. imported from PROfit). The first implementation could randomly allocate a different set of eyes to each population face. When a good match had been identified, the relevant set of eyes could be fixed and evolution continued with that choice. Note that Caldwell & Johnston (1991) demonstrate an advantage for “freezing” facial features in this way in an evolutionary context. A later software version might arrange the eyes to fit along a number of psychological dimensions, enabling similar sets of eyes to be displayed on request. For example, one dimension may be eye colour and would enable only light coloured eyes to be presented. Other dimensions could be derived via mathematical scaling techniques, such as multidimensional scaling (Kruskal, 1964; and Kruskal & Wish, 1978), known to be successful in categorizing face stimuli (e.g. Johnston, Milne, Williams & Hosie, 1997; Shepherd, Ellis & Davies, 1977; Valentine, 1991; and Vokey & Read, 1992)

Further development could also be valuable to correct hairstyles that appear unnatural when fitted to the head; the effect is best described as “floating” hair, an example of which can be seen in Figure 83. The problem has been resolved to date by the manual blending of the outline of the hair over the forehead in Photoshop. Theoretically, this process can be automated when an image is imported from the external photofit system (e.g. PROfit). This might involve the creation of a “blending mask” that enables fading of pixels between the selected hairstyle and the external features, rather like the mask used to fuse the internal and external features when currently generating a face (Figure 18). The blending mask could be created from a simple algorithm that searched for the outline of the hair.

Multiple Witnesses

One method of further enhancing performance could arise through the use of multiple witnesses. It is sometimes the case that there are multiple witness to a crime. When this occurs, the police may assign different tasks to different witnesses: one witness may create a composite, another may select photographs from a mugshot album, and a third may be used to chose a suspect from a line-up. It is clear that although the assigned tasks are rather different, information from multiple witnesses could be valuable in creating a photofit (e.g. Bruce, Ness, Hancock, Newman & Rarity, submitted; and McNeil, Wray, Hibler, Foster, Rhyne & Thibault, 1987). McNeil et al. made a “modal” composite from the highest selected facial features used in

32 Identikits. They found that modal composites were rated significantly higher than the constituent Identikits. In contrast, Bruce, Ness, Hancock, Newman & Rarity found that simultaneously presenting composites from 4 people resulted in a 16% increase in identification compared with using a single composite for recognition.

The proposed study would engage a number of subjects evolving a common target face. All subjects would see the same sets of faces but their selections would be used to weight the faces that are selected as parents (i.e. to modulate the fitness function). As mentioned previously, there is evidence of considerable overlap between the selections made by individuals. This said, the most frequently selected “best” face in that pilot study was only chosen about half the time (53%) and suggests that input from multiple witnesses could be valuable in potentially tuning the fitness value of the selected faces. Like the creation of a $\frac{3}{4}$ view database mentioned above, this research is also planned in the next 2 years as part of the CRIME-VU⁷² project.

Final Comments

Further research and development is clearly necessary before a photofit product becomes commercially available. Arguably, the most pressing area of research concerns the creation of EvoFITs from memory. Regarding development, the most important issue concerns the anonymity of the images that form the database. The resolution of this issue is viewed as a prerequisite for product adoption in forensic circles. A related issue of course is how the system would perform in conjunction with existing operating procedures such as the eliciting of a verbal description. Then there is the issue concerning the utility of colour, a multiple witness mode of construction and the use of $\frac{3}{4}$ view representations. In addition, one hopes to be able to examine EvoFIT performance should a conviction result from Operation Mallard. Such a case study may shed light into performance tweaks that could lead to better performance in future. Of course, one would also welcome the opportunity to apply the system to other criminal cases with the same rationale in mind.

Much of the aforementioned work concerning the EvoFIT system is planned to take place over the next 2 years by myself and a further research assistant. To date, it has been shown that the general holistic/evolutionary approach is promising but continued work should be performed *before* any adoption in forensic circles. This is believed to be important to avoid a product being released and then found to be of limited value later, as has been the fate of photofit systems to date.

Glossary of Abbreviations

ACPO(S)	Association of Chief Police Officers (Scotland)
AFSS	Anchored Face Similarity Scale
CADC	Computer-Aided Design Centre
CHR	Conditional Hit Rate
CI	Cognitive Interview
CMR	Correlation between MSE and Rating scores
DF	Degrees of Freedom
EFI	External Feature Image
EFIT	Electronic Facial Identification Technique
EvoFIT	A Holistic, Evolutionary Facial Imaging System
FNP	Facial Normal Palette
FSP	Facial Shape Palette
FSPBT	Facial Shape Palette with the Best Texture (from the previous generation)
FTP	Facial Texture Palette
FTPBS	Facial Texture Palette with the Best Shape (from the previous generation)
GA(s)	Genetic Algorithm(s)
IFI	Internal Feature Image
JND	Just Noticeable Difference
MAMP	Mac-a-mug Pro
MDSS	Multidimensional Similarity Space
MSE	Mean Square Error
PC(s)	Principle Component(s)
PCA	Principal Components Analysis
RMS	Root Mean Square (error measure)
SD	Standard Deviation

References

- ACPO(S) (2000). National Working Practices in Facial Imaging. ACPOS Working Group.
- 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*, 14-21, July 1987.
- Baker, E., & Selzer, M. (1994). Evolving line drawings. *Graphics Interface Proceedings*, Banff, Canada, May 1994.
- Baker, E., & Selzer, M. (1998). The mug-shot search problem, *Visual Interface '98 Proceedings*, Vancouver, BC, Canada, June 1998.
- Bennett, P. (1986). Face recall: A police perspective. *Human Learning*, 5, 197-202.
- Bennett, P. (2000). The use of multiple composites in suspect identification. Proceedings of the 3rd UK national conference on cranio-facial identification. Manchester, May 2000.
- Berman, G.L., & Cutler, B.L. (1998). The influence of processing instructions at encoding and retrieval on face recognition accuracy. *Psychology, Crime & Law*, 4, 89-106.
- BioVirtual (2001). 3DMeNow by BioVirtual. <http://www.biovirtual.com>.
- Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces, *SIGGRAPH'99 Conference Proceedings*.
- Bower, G. H., & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology*, 103, 751-757.
- Brace, N., Pike, G., & Kemp, R. (2000). Investigating E-FIT using famous faces. In A. Czerederecka, T. Jaskiewicz-Obydzinska & J. Wojcikiewicz (Eds.). *Forensic Psychology and Law*. Krakow: Institute of Forensic Research Publishers.
- Brennan, S.E. (1985). The caricature generator. *Leonardo*. 18, 170-178.

- Brown, G.D.A., Hulme, C., Hyland, P.D., & Mitchell, I.J. (1994). Cell suicide in the developing nervous systems: a functional neural network model. *Cognitive Brain Research*, 2, 71-75.
- Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition, *British Journal of Psychology*, 73, 105-116.
- Bruce, V. (1986). Influences of familiarity on the processing of faces. *Perception*, 15, 387-397.
- Bruce, V. (1995). Perceiving and recognizing faces. In I. Roth & V. Bruce (Eds.). *Perception and representation* (2nd ed.). Open University Course Unit D309. Part III.
- Bruce, V. (1988). *Recognising faces*. LEA.
- Bruce, V., Burton, A.M., & Dench, N. (1994). What's distinctive about a distinctive face? *Quarterly Journal of Experimental Psychology*, 47A, 119-149.
- Bruce, V., Doyle, T., Dench, N., & Burton, A.M. (1991). Remembering facial configurations. *Cognition*, 38, 109-144.
- Bruce, V., Hanna, E., Dench, N., Healey, P. & Burton, M. (1992). The importance of "mass" in line-drawings of faces. *Applied Cognitive Psychology*, 6, 619- 628.
- Bruce, V., Healey, P., Burton, A.M., Doyle, T., Coombes, A., & Linney, A. (1991). Recognising facial surfaces. *Perception*, 20, 755-769.
- Bruce, V., Ness, H., Hancock, P.J.B., Newman, C., & Rarity, J. (submitted). Four heads are better than one. Combining face composites yields improvements in face likeness. Submitted to the *Journal of Applied Psychology*.
- Bruce, V., Valentine, T., & Baddeley, A.D. (1987). The basis of the 3/4 view advantage in face recognition. *Applied Cognitive Psychology*, 1, 109-120.
- Bruce, V., & Young, A.W. (1986). Understanding face recognition, *British Journal of Psychology*, 77, 305-327.
- Bruce, V., & Young, A.W. (1998). *In the eye of the beholder*. OUP.

- Bruck, M., Cavanagh, P., & Ceci, S.J. (1991). Fortysomething: recognizing faces at one's 25th reunion. *Memory & Cognition*, 19 (3), 221-228.
- Brunelli, R., & Mich, O. (1995). SpotIt! An interactive Identikit System, IRST Tech. Rep. #9507-03.
- Brunelli, R., & Poggio, T. (1993). Face Recognition: Features versus Templates, *IEEE Transactions on PAMI*, 15(10), 1042-1052.
- Burton, A.M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor quality video: evidence from security surveillance. *Psychological Science*, 10 (3), 243-248.
- Cabeza, R., Bruce, V., Kato, T., & Oda, M. (1999). The prototype effect in face recognition: extension and limits. *Memory and Cognition*, 27(1), 139-151.
- Cabeza, R., & Kato, T. (2000). Features are also important: Contributions of featural and configural processing to face recognition. *Psychological Science*, 11(5), 429-433.
- Caldwell, C., & Johnston, V.S. (1991). Tracking a Criminal Suspect Through "Face-Space" with a Genetic Algorithm, *Proceedings of the Fourth International Conference on Genetic Algorithms*, 416-421. Morgan Kaufmann Publishers.
- Christie, D., Davies, G.M., Shepherd, J.W., & Ellis, H.D. (1981). Evaluating a new computer-based system for face recall, *Law and Human Behaviour*, 2/3, 209-218.
- Christie, D., & Ellis, H. (1981). Photofit constructions versus verbal descriptions, *Journal of Applied Psychology*, 66, 358-363.
- Clifford, B.R., & Davies, G.M. (1989). Procedures for obtaining identification evidence. In Rankin D.C., *Psychological methods in criminal investigation and evidence*, (47-96). New York: Springer.
- Cohen, M.E., & Nodine, C.F. (1978). Memory processes in facial recognition and recall. *Bulletin of the Psychonomic Society*, 12 (4), 317-319.
- Comish, S. (1987). Recognition of facial stimuli following an intervening task involving the Identikit. *Journal of Applied Psychology*, 72, 488-491.

- Cootes, T.F., Walker, K.N., & Taylor, C.J. (2000). View-Based Active Appearance Models. *Proceedings of the International Conference on Face and Gesture Recognition*. 227-232.
- Courtois, M.R., & Mueller, J.H. (1981). Target and distractor typicality in facial recognition. *Journal of Applied Psychology*, 66, 639-645.
- Craw, I., & Cameron, P. (1991). Parameterising images for recognition and reconstruction. *Proceedings of the British Machine Vision Conference BMCV '91*, Turing Institute Press and Springer Verlag.
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). Improving the reliability of eyewitness identifications: Putting context into context. *Journal of Applied Psychology*, 72, 629-637.
- Cutshall, J.L., & Yuille, J.C. (1989). Field studies of eyewitness memory of actual crime scenes. In Raskin, D.C. (Ed.), *Psychological methods in criminal investigation and evidence* (97-124). New York: Springer.
- Davies, G.M. (1978). Face recognition: issues and theories. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.). *Practical aspects of memory*. New York: Academic Press.
- Davies, G.M. (1982). Representing the appearance of a 'wanted' suspect: realism vs. simplicity. Paper presented at the *British Psychological Society (Welsh Branch) Conference on Law and Psychology*, Swansea, July 19-23, 1982.
- Davies, G.M. (1983a). Forensic face recall: the role of visual and verbal information. In S.M.A. Lloyd-Bostock & B.R. Clifford (Eds.). *Evaluating witness evidence*. London: John Wiley & Sons Ltd., Chapter 6.
- Davies, G.M. (1983b). The recognition of persons from drawings and photographs. *Human Learning*, 2, 237-249.
- 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., G., & Shepherd, J. (1978). Face identification: The influence of delay upon accuracy of photofit construction, *Journal of Police Science and Administration*, 6(1), 35-42.

Davies, G. M., & Milne, A. (1982). Recognizing faces in and out of context. *Current Psychological Research*, 2(4), 235-246.

Davies, G. M., & Milne, A. (1985). Eyewitness composite production: a function of physical reinstatement of context. *Criminal Justice and Behavior*, 12, 209-220.

Davies, G. M., Milne, A., & Shepherd, J. (1983). Searching for operator skills in face composite reproduction. *Journal of Police Science and Administration*, 11(4), 405-9.

Davies, G.M., & Oldman, H. (1999). The impact of character attribution on composite production: A real world effect? *Current Psychology: Developmental, Learning, Personality, Social*. 18 (1), 128-139.

Davies, G. M., Shepherd, J., & Ellis, H. (1977). Similarity effects in face recognition. *American Journal of Psychology*, 92 (3), 507-523.

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

Davies, G. M., Shepherd, J., Shepherd, J., Flin, R., & Ellis, H. (1986). Training skills in police photofit operators, *Policing*, 2, 35-46.

Davies, G.M., & Thasen, S. (2000). Closed-circuit television: How effective an identification aid? *British Journal of Psychology*, 91(3): 411-426.

Davies, G. M., van der Willik, P., & Morrison, L.J. (2000). Facial Composite Production: A Comparison of Mechanical and Computer-Driven Systems, *Journal of Applied Psychology*, 85, 1, 119-124.

Dawkins, R. (1991). *The blind watchmaker*. London: Penguin, 1991.

Dodson, C.S., Johnson, M.K., & Schooler, J.W. (1997). The verbal overshadowing effect: Source confusion or strategy shift? *Memory & Cognition*, 25, 129-139.

Dukes, W.F., & Bevan, W. (1967). Stimulus variation and repetition in acquisition of naming responses. *Journal of Experimental Psychology*, 74, 178-181.

Ellis, H.D. (1975). Recognising faces. *British Journal of Psychology*, 66, 404-426.

- Ellis, H.D. (1984). Practice aspects of face memory. In G. R. Wells & E. F. Loftus (Eds.), *Eyewitness Testimony. Psychological Perspectives*. Cambridge, England: Cambridge University Press, 1984. Ch. 2.
- Ellis, H. D. (1986). Face recall: A psychological perspective, *Human Learning*, 5, 1-8.
- Ellis, H., Davies, G. M., & Shepherd, J. (1978a). A critical examination of the photofit system for recalling faces, *Ergonomics*, 21, 4, 297-307.
- Ellis, H., Davies, G. M., & Shepherd, J. (1978b). Remembering pictures of real and 'unreal' faces: some practical and theoretical considerations, *British Journal of Psychology*, 69, 467-474.
- Ellis, H.D., & Shepherd, J.W. (1992). Face memory - theory and practice. In M.M. Gruneberg & P.E. Morris (Eds.). *Aspects of Memory, Vol. 1, The practical aspects* (2nd ed.). London: Routledge.
- Ellis, H. Shepherd, J., & Davies, G. M. (1975). Use of photo-fit for recalling faces, *British Journal of Psychology*, 66, 29-37.
- Ellis, H., Shepherd, J., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition, *Perception*, 8, 431-439.
- Ellis, H.D., Shepherd, J.W., & Davies, G.M. (1980). The deterioration of verbal descriptions of faces over different delay intervals. *Journal of Police Science and Administration*, 8, 101-106.
- Fallshore, M., & Schooler, J.W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 6, 1608-1623.
- FIC (1999). Crime management division: facial imaging course. Scottish Police College.
- Finger & Pezdek (1999). The effect of the cognitive interview on face identification accuracy: release from verbal overshadowing. *Journal of Applied Psychology*, 84, 3, 340-348.
- Flin, R., Markham, R., & Davies, G. M. (1990). Making faces: Developmental trends in the construction and recognition of photofit face composites. *Journal of Applied Developmental Psychology*, 10(2), 131-145.

- Gibling, F., & Bennett, P. (1994). Artistic enhancement in the production of photofit likeness: an examination of its effectiveness in leading to suspect identification, *Psychology, Crime & Law*, 1, 93-100.
- Gillenson, M., & Chandrasekaran, B. (1975). A heuristic strategy for developing human facial images on a CRT, *Pattern Recognition*, 7, 187-196.
- Goffredson, M.R., & Polakowski, M. (1995). Information retrieval: reconstructing faces. In N. Brewer & C. Wilson (Eds.). *Psychology & Policing*. Hillsdale, NJ: Lawrence Erlbaum. 101-117.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.
- Goldstein, A.G., & Chance, J.E. (1979). Visual recognition memory for complex configurations. *Perception and Psychophysics*, 9, 237-241.
- Goldstein, A.G., Stephenson, B., & Chance, J. (1977). Face recognition memory: distribution of false alarms. *Bulletin of the Psychonomic Society*, 9 (6), 416-418.
- Green, D. L., & Geiselman, R. E. (1989). Building composite facial images: Effects of feature saliency and delay of construction. *Journal of Applied Psychology*, 74, 714-721.
- Haig, N.D. (1984). The effect of feature displacement on face recognition. *Perception*, 13, 505-512.
- Hall, D.F. (1976). Obtaining eyewitness identification in criminal investigations: applications of social and experimental psychology (Doctoral dissertation, Ohio State University, Columbus). *Dissertation Abstracts International*, 37, 2569B.
- Hancock, P.J.B. (2000). Evolving faces from principal components. *Behavior Research Methods, Instruments and Computers*, 32-2, 327-333.
- Hancock, P.J.B., Bruce, V., & Burton, A.M. (1997). Testing principal component representations for faces. In J.A. Bullinaria, D.W. Glasspool & G. Houghton (Eds.). *Proceedings of 4th Neural Computation and Psychology Workshop*, London: Springer-Verlag, 84-97.
- Hancock, P.J.B., Bruce, V., & Burton, A.M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences*, 4-9, 330-337.

- Hancock, P.J.B., Burton, A.M., & Bruce, V. (1996). Face processing: human perception and principal components analysis. *Memory & Cognition*, 24, 26-40.
- Hancock, P.J.B., & Frowd, C.D. (1999). Evolutionary generation of faces. *Proceedings AISB99*, Edinburgh, 6 - 9th April 1999.
- Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces, *Journal of Experimental Psychology: Human Perception and Performance*, 22 (4), 986-1004.
- Hill, H., Bruce, V., & Akamatsu, S. (1995). Perceiving the sex and race of faces: the role of shape and colour. *Proceedings of the Royal Society*. 367-373.
- Homa, D. Haver, B., & Schwartz, T. (1976). Perceptibility of schematic face stimuli: Evidence for a perceptual Gestalt. *Memory & Cognition*, 4, 176-185.
- Honeyman, H.B. (unpublished data). A psychophysical investigation into the observable relational changes of four feature changes. *Psychology Dissertation*, University of Stirling, 2001.
- ImageWare (2000). ImageWare Systems, Inc. <http://www.iwsinc.com/>
- Inn, D., Walden, K.J., & Solso, R.L. (1993). Facial prototype formation in children. *Bulletin of the Psychonomic Society*, 31(3), 197-200.
- Johnston, R.A., Milne, A.B., Williams, C., & Hosie, J. (1997). Do distinctive faces come from outer space? An investigation of the status of Multidimensional Face-Space. *Visual Cognition*, 4(1), 59-67.
- Kemp, R., Pike, G., White, P., & Musselman, A. (1996). Perception and recognition of normal and negative faces: the role of shape from shading and pigmentation cues. *Perception*, 25(1), 37-52.
- Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: photographs, credit cards and fraud. *Applied Cognitive Psychology*, 21, 297-307.
- King, D. (1971). The use of photofit 1970-1971: a progress report. *Police Research Bulletin*, 18, 40-44.

- Kirby, M., & Sirovich, L. (1990). Application of the karhunen-loeve procedure for characterization of human faces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1), 103-108.
- Koehn, C.E. (1996). Theoretical and applied evaluations of facial composite systems. *Dissertation Abstracts International, B, The Sciences and Engineering*, 56(8-B), 4640.
- Koehn, C.E., & Fisher R.P. (1997). Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime & Law*, 3, 215-224.
- Kovera, M.B., Penrod, S.D., Pappas, C., & Thill, D.L. (1997). Identification of computer generated facial composites, *Journal of Applied Psychology*, 82(2), 235-246.
- Krouse, F.L. (1981). Effects of pose, pose change, and delay on face recognition performance. *Journal of Applied Psychology*, 66(5), 651-654.
- Kruskal, J.B. (1964). Multidimensional scaling: A numerical method. *Psychometrika*, 29, 115-129.
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage University Series.
- Lander, K. (unpublished). The role of dynamic information in the recognition of famous faces. Unpublished Ph.D. thesis. University of Stirling, August 1999.
- Laughery, K.R., Duval, C., & Wogalter, M.S. (1986). Dynamics of facial recall. In Ellis, H.D., Jeeves, M.A., Newcombe, F., & Young, A. (Eds.). *Aspects of face processing*. Dordrecht: Martinus Nijhoff. 373-387.
- 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.
- Laughery, K.R., & Fowler, R. (1980). Sketch artist and identikit procedures for generating facial images, *Journal of Applied Psychology*, 65, 307-316.
- Laughery, K.R., & Smith, V.L. (1978). Suspect identification following exposure to sketches and identikit composites. *Proceedings of the Human Factors Society 22nd Annual Meeting*, Detroit.

- Le Cun, Y, Denker, J.S., & Solla, S.A. (1990). Optimal brain damage. *Advances in Neural Information Systems*, 2, 598-605.
- Leder, H. (1996). Line drawings of faces reduce configural processing. *Perception*, 25, 355-366.
- Levi, A.M., Jungman, N., Ginton, A., Aperman, A., & Noble, G. (1995). Using similarity judgments to conduct a mugshot album search. *Law and Human Behavior*, 19, 649-661.
- Light, L.L., Kayra-Stuart, F., & Hollander, S., (1979). Recognition memory for typical and unusual faces, *Journal of Experimental Psychology: Human Learning and Memory*, 5(3), 212-228.
- Logie, R.H., Baddeley, A.D., & Woodhead, M.M. (1987). Face recognition, pose and ecological validity. *Applied Cognitive Psychology*, 1, 53-69.
- Malpass, R.S. (1996). Enhancing eyewitness memory. In S.L. Sporer, R. S., Malpass, & G. Koehnken (Eds.). *Psychological issues in eyewitness identification*. Hillsdale, NJ: Lawrence Erlbaum. Ch. 8.
- Malpass, R.S., & Devine, P.G. (1981). Eyewitness identification: lineup instructions and the absence of the offender. *Journal of Applied Psychology*, 66, 4, 482-489.
- Matlab (1997). Using matlab. Mathworks Inc.
- Mauldin, M., & Laughery, K. (1981). Composite production effects upon subsequent facial recognition, *Journal of Applied Psychology*, 66, 351-357.
- McKelvie, S.J. (1988). The role of spectacles in facial memory: a replication and extension. *Perceptual & Motor Skills*, 66, 651-658.
- McNeil, J.E., Wray, J.L., Hibler, N.S., Foster, W.D., Rhyne, C.E., & Thibault, R. (1987). Hypnosis and Identi-kit: a study to determine the effect of using hypnosis in conjunction with the making of identikit composites. *Journal of Police Science and Administration*, 15, 63-67.
- Meissner, C.A., & Brigham, J.C. (2001). A meta-analysis of the verbal overshadowing effect in face identification, *Applied Cognitive Psychology*, 15(6), 603-616.
- Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge: MIT Press.

- Mozer, M.C., & Smolensky, P. (1989). Using relevance to reduce network size automatically. *Connection Science*, 1(1), 1-16.
- O'Donnell, C., & Bruce, V. (in press). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception*.
- O'Toole, A.J., Abdi, H., Deffenbacher, K.A., & Valentin, D. (1993). Low dimensional representation of faces in high dimensions of the space. *Journal of the Optical Society of America A*, 10,405-410.
- O'Toole, A.J., Deffenbacher, K.A., Valentin, D., & Abdi, H. (1993). Structural aspects of face recognition and the other race effect. *Memory & Cognition*, 22 (2), 208-224.
- O'Toole, A. J., Edelman, S., & Buelthoff (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, 38, 2351-2363.
- Patterson, K.E., & Baddeley, A.D. (1977). When recognition fails. *Journal of Applied Psychology*, 3(4), 406-417.
- Penry, J. (1970). Photofit. *Police Journal*, 302-316.
- Penry, J. (1974). Photo-Fit. *Forensic Photography*, 3(7), 4-10.
- Perrett, D., Benson, P.J., Hietanen, J.K., Oram, M.W., & Dittrich, W.H. (1995). When is a face not a face? In R. Gregory, J. Harris, P. Heard & D. Rose (Eds.). *The Artful Eye*, Oxford University Press.
- Rakover, S.S., & Cahlon, B. (1996). To catch a thief with a recognition test: the model and some empirical results. *Cognitive Psychology*, 21, 423-468.
- Rattner, A. (1988). Convicted but innocent: wrongful conviction and the criminal justice system. *Law and Human Behavior*, 12, 283-293.
- Read, J.D. (1979). Rehearsal and recognition of faces. *American Journal of Psychology*, 92, 71-85.
- Rhodes, G., Brennan, S.E., & Carey, S. (1987). Identification and ratings of caricatures: implications for mental representations of faces. *Cognitive Psychology*, 19, 473-494.

- Rumelhart, D.E., & McClelland, J.L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*. Vol. 1: Foundations, Cambridge: MIT.
- Schooler, J.W., & Engstler-Schooler, T.Y. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology*, 22, 36-71.
- Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122 (2): 166-183.
- Shapiro, P. N., & Penrod, S.D. (1986). Meta-analysis of facial identification rates. *Psychological Bulletin*, 100, 139-156.
- Shavelson, R.J. (1981). *Statistical reasoning for the behavioral sciences*. Allyn & Bacon.
- Shepherd, J.W. (1986). An interactive computer system for retrieving faces. In Ellis, H.D., Jeeves, M.A., Newcombe, F., & Young, A. (Eds.). *Aspects of face processing*. Dordrecht: Martinus Nijhoff.
- Shepherd, J., Davies, G. M., & Ellis, H.D. (1978). How best shall a face be described? In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.). *Practical aspects of memory*. New York: Academic Press.
- Shepherd, J., & Ellis, H. (1996). Face recall - methods and problems. In Sporer, S.L., Malpass, R.S., & Koehnken, G. (Eds.). *Psychological issues in eyewitness identification*. Hillsdale, NJ: Lawrence Erlbaum. Ch. 5.
- Shepherd, J.W., Ellis, H.D., & Davies, G.M. (1977). Perceiving and remembering faces. *Home Office report POL/72/1675/24/1*.
- Shepherd, J.W., Ellis, H.D., & Davies, G.M. (1982). *Identification evidence: a psychological evaluation*. Aberdeen: Aberdeen University Press.
- Shepherd, J.W., Ellis, H.D., McMurrin, M., & Davies, G.M. (1978). Effect of character attribution on Photofit construction of a face. *European Journal of Social Psychology*, 8, 263-8.
- Shepherd, F., Gibling, H.D., Ellis, H.D. (1991). The Effects of distinctiveness, presentation time and delay on face recognition. *European Journal of Cognitive psychology*, 3, 137-45.

Sirchie (2000). Sirchie Finger Print Laboratories Inc. <http://www.sirchie.com/>

Visatex (2000). Digital Descriptor Systems. <http://www.ddsi-cpc.com/>

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces, *J. Opt. Soc. Amer. A*, 4, 519-524.

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

Sporer, S.L. (1988). Long-term improvement of facial recognition through visual rehearsal. In M. Gruneberg, P. Morris, & R. Sykes (Eds.). *Practical aspects of memory: Current research and issues*. Vol. 1. Chichester: Wiley. 182-188.

Sporer, S.L., Koehnken, G., & Malpass, R.S. (1996). Introduction: 200 years of mistaken identification. In S.L. Sporer, R. S., Malpass, & G. Koehnken (Eds.). *Psychological issues in eyewitness identification*. Hillsdale, NJ: Lawrence Erlbaum. Ch. 1.

Sporer, S.L., Penrod, S.L., Read, J.D., & Cutler, B. (1995). Choosing, confidence and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118, 315-327.

Tanaka, J.W., & Farah, M.J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 46A, 2, 225-245.

Troje, N.F., & Vetter, T. (1996). Representation of human faces. *Technical report*, Max-Planck-Institut, Tubingen, Germany.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology*, 43A, 161-204.

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, 15, 525-536.

Valentine, T. & Endo, M. (1992). Towards an exemplar model of face processing: the effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology*, 44A, 671-703.

Vokey, J.R., & Read, J.D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, 20(3), 291-302.

Wagenaar, W., & Van der schrier, J.H. (1996). Face recognition as a function of distance and illumination: a practical tool for use in the courtroom. *Psychology, Crime & Law*, 2, 321-332.

Walker-Smith, G. J. (1978). The effects of delay and exposure duration in a face recognition task, *Perception & Psychophysics*, 24 (1), 63-70.

Wells, G.L., & Hryciw, B. (1984). *Memory & Cognition*, 12, 4, 338-344.

Wells, G. L., & Turtle, J.W. (1988). What is the best way to encode faces? In M.M. Gruneberg, P. Morris, & R. Sykes (Eds.). *Practical aspects of memory: Current research and issues*. Vol. 1. Chichester: Wiley. 163-168.

Westerman, D. L., & Larsen, J. D. (1997). The verbal overshadowing effect: Evidence for a general shift in processing. *American Journal of Psychology*, 110, 417-428.

Wogalter, M., & Marwitz, D. (1991). Face composite construction: In view and from memory quality improvement with practice, *Ergonomics*, 22, 333-343.

Yin, R.K. (1969). Looking at upside down faces. *Journal of Experimental Psychology*, 81, 141-5.

Young, A.W., Hay, D.C., McWeeny, K.H., Flude, B.M. & Ellis, A.W. (1985). Matching familiar and unfamiliar faces on internal and external features, *Perception*, 14, 737-746.

Young, A.W., Hellawell, D. & Hay, D.C. (1987). Configural information in face perception, *Perception*, 16, pp. 747-759.

Zavala, R.T. (1972). Determination of facial features used in identification. In A. Zavala, J.J. Paley, & R.R.J. Gallati (Eds.). *Personal appearance identification*. Chapter VI. Charles C Thomas.

Zeda (1998). *CD-Fit user guide*. Zeda ABM Ltd.

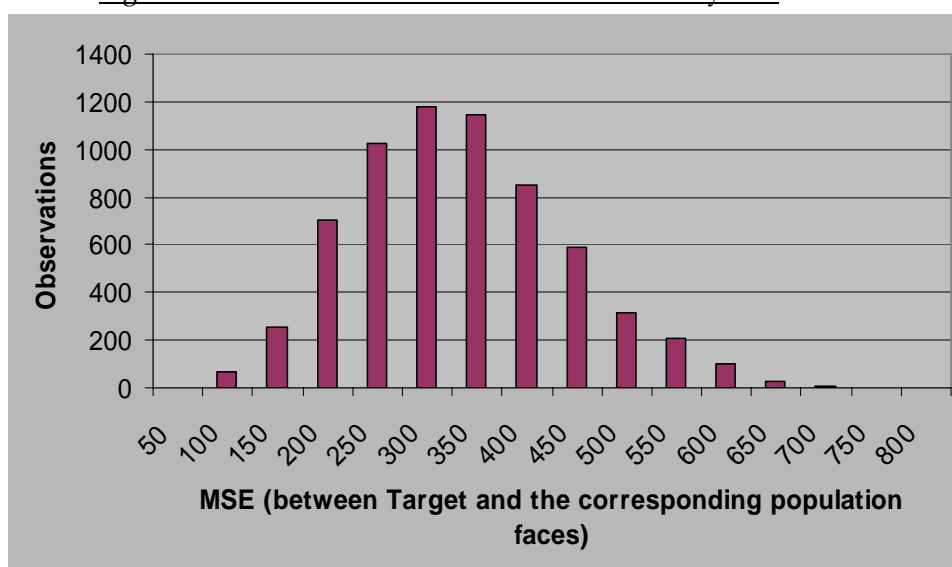
Appendices

Appendix A: Mean Square Error Measures of Facial Images

Mean Square Error (MSE) is a measure of the average error between two vectors (or matrices). It is bounded by zero at the lower limit, occurring when the two vectors are equal. The upper limit is specified by the square of the maximum possible difference along any dimension between the vectors. For the 8-bit grey scale images used in this study, each pixel has a range of 0 to 255, resulting in an upper limit of $(0-255)^2$ or 65025. Note that this limit is with one image pure white and the other, pure black. For facial images, this does not take into account that generated images will be neither black nor white and contain considerable structure.

An attempt was made to gain an estimate of the upper limit for MSE in the Pilot Study (Chapter 2). This was achieved by computing the MSE scores between the five targets and all the population faces generated in Chapter 2. The distribution of scores is shown below –

Figure 84: Distribution of MSE Scores for the Pilot System



It can be seen that scores are roughly normally distributed (though a slight skew to the left is apparent). The mean is 308.6 (SD 106.5). Importantly, the graph indicates that the maximum MSE found is in the 750 category; this observation had a value of 735.5 and for this data set indicates the limit for MSE in the Pilot.

In general, there is concern by the author regarding the utility of the MSE as a measure of performance. At the lower end, this appears to be appropriate (since it produces a test for identity) but the question arises as to the relationship between the MSE and the *perceptual* similarity of a face to the target. In a small study, attempting to address this issue, 6 faces were

selected from a pool of a 100 randomly generated images (also taken from the Pilot Study) such that images were spaced apart by an MSE of approximately 50 pixels⁷⁴. 20 Subjects were told to rank order the images into similarity with the target face. A correlation of the average ranking resulted in a near perfect relationship ($r=0.91$). This indicates that MSE *can* be considered a sensible measure of facial similarity.

⁷⁴ The MSE scores from the given target were 99, 152, 213, 256, 304, 355 and 399.

Appendix B: Famous Face Stimuli Used for Experiment 10

Appendix C: EFIT Description Sheet Used for Experiment 10

E-Fit Description

EiFit No.:-

CF No. :-

Date :-

Time :-

OIC :-

Witness Details:-

General Description and Events;-

Face in Detail

Shape

Hair

Eyebrows

Eyes

Nose

Mouth

Ears

%Likeness




Witness Signature

E-Fit Operator:-

Appendix D: Composites Created in Experiment 10

The following are the photofits of famous faces created in Experiment 10 from (a) the EFIT system and (b) the EvoFIT system.

(a) Composites Created by EFIT

Celebrity	EFIT
Terry Wogan	
Bruce Willis	
Bob Geldof	

Tony Blair



David Bowie



Jimmy Nail



Jim Carrey



Nicholas Lyndhurst



Woody Allen

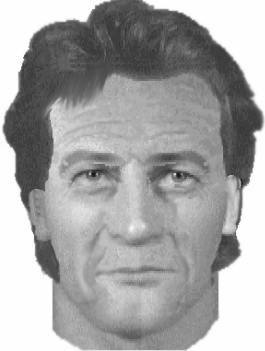





John Travolta



Tom Hanks



<p>Mel Gibson</p>	
<p>Jonathan Ross</p>	
<p>Al Pacino</p>	
<p>Richard Gere</p>	

Robert De Nero



Tom Cruise








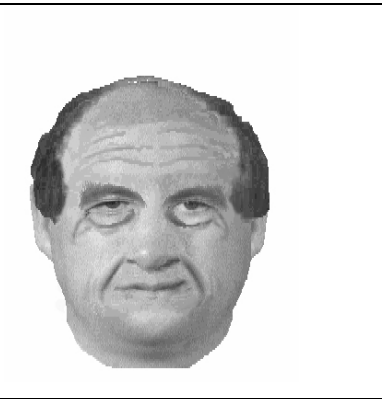


Cliff Richard



Dale Winton



<p>Michael Douglas</p>	
<p>Mick Jagger</p>	
<p>Robin Williams</p>	
<p>Ronan Keating</p>	

<p>Michael Owen</p>	
<p>Danny DeVito</p>	
<p>George Clooney</p>	
<p>Graham Norton</p>	

Brad Pitt




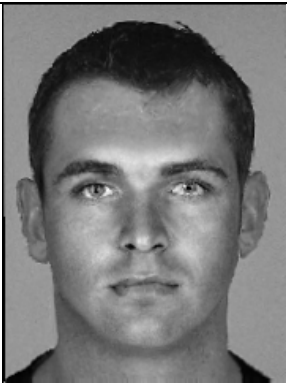
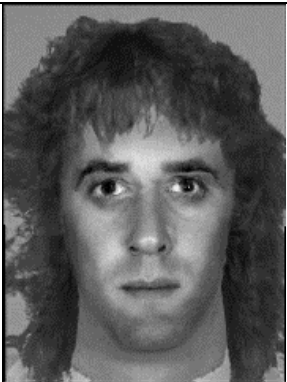
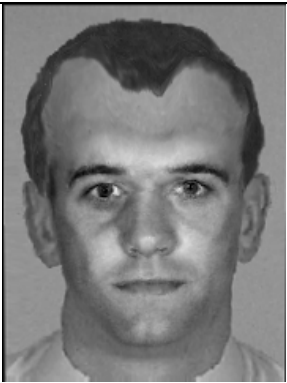
Michael Barrymore

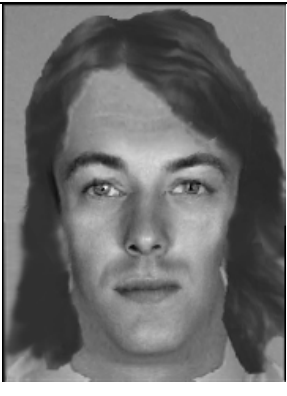
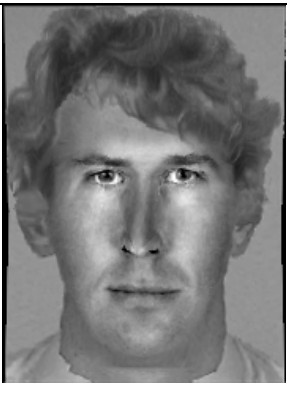
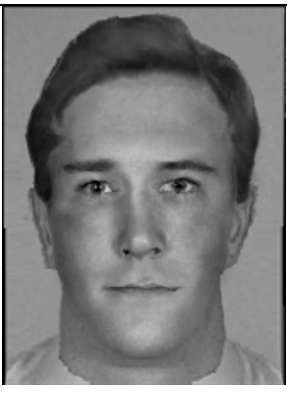
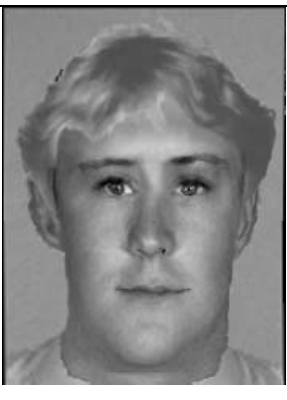



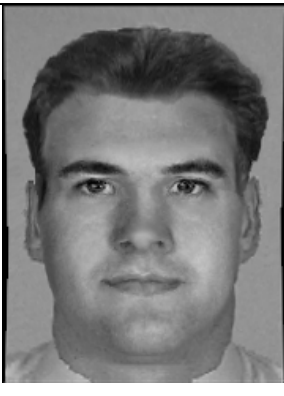

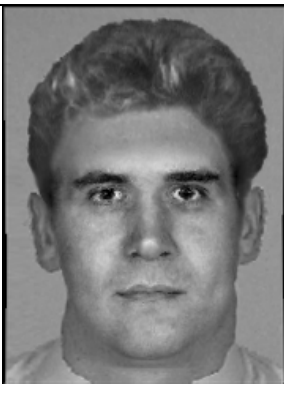
Michael Caine

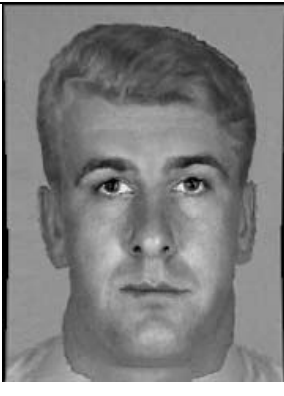
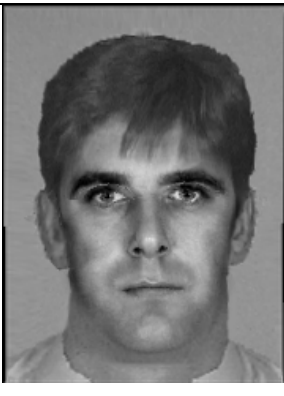
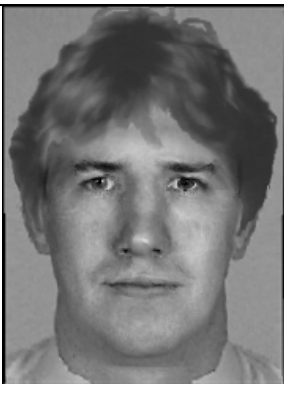




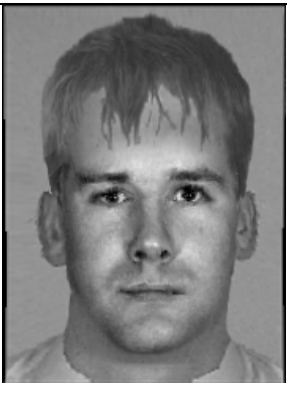
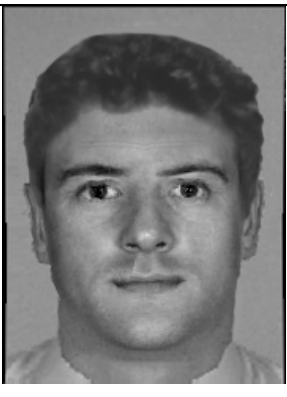
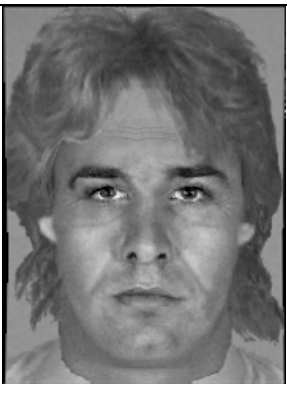
(b) Composites Created by EvoFIT



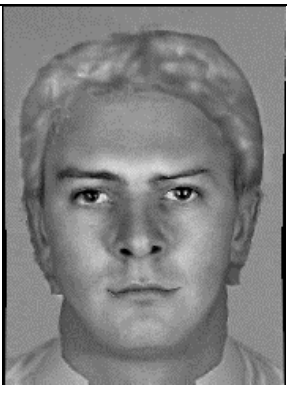
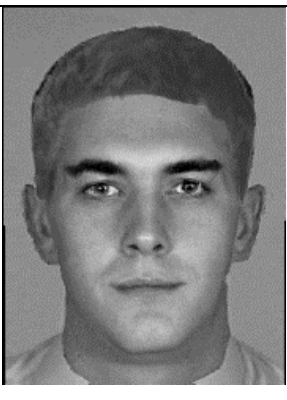
Celebrity	EvoFIT	
Terry Wogan		
Bruce Willis		
Bob Geldof		
Tony Blair		

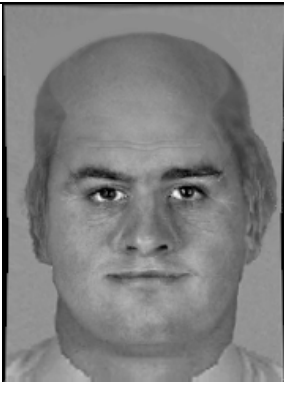
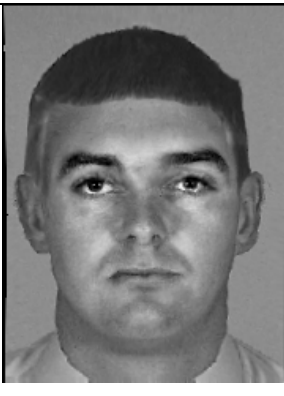


<p>David Bowie</p>			
<p>Jimmy Nail</p>			
<p>Jim Carrey</p>			
<p>Nicholas Lyndhurst</p>			



<p>Woody Allen</p>			
<p>John Travolta</p>			
<p>Tom Hanks</p>			
<p>Mel Gibson</p>			

<p>Jonathan Ross</p>			
<p>Al Pacino</p>			
<p>Richard Gere</p>			
<p>Robert De Nero</p>			

<p>Tom Cruise</p>			
<p>Cliff Richard</p>			
<p>Dale Winton</p>			
<p>Michael Douglas</p>			

<p>Mick Jagger</p>			
<p>Robin Williams</p>			
<p>Ronan Keating</p>			
<p>Michael Owen</p>			

<p>Danny DeVito</p>			
<p>George Clooney</p>			
<p>Graham Norton</p>			
<p>Brad Pitt</p>			

<p>Michael Barrymore</p>			
<p>Michael Caine</p>			

Appendix E: Targets used for Experiment 11

Appendix F: EvoFIT Operating Procedures for Experiment 11

EvoFIT: A Holistic, Evolutionary Facial Imaging System

Operator's Manual

*Face Perception Group
University of Stirling
Stirling*

Creating a Composite: A Quick Glance

The recommended operating sequence to create a photofit with this system is -

1. Start PROfit and Photoshop
2. Start a new photofit session (run the EvoFIT program then select *Create ...* from the main menu)
3. Selection of hairstyle (in PROfit) & modification (in Photoshop)
4. Shape selection (at least 6 faces over at least 3 screens)
5. Texture selection (at least 6 faces over at least 3 screens)
6. Selection of *best face*
7. *Make Next Generation* (to create offspring faces)
8. Selection of Shapes, Textures and *best face* as before
9. Modify *best face* in *Feature Shifter* (right-click on best face in *Normal* display mode)
10. *Make Next Generation* (to create more offspring faces)
11. Repeat from 8 until acceptable likeness achieved
12. Select desired photofit and save to disk (*Save As* under the EvoFIT menu item)
13. Exit (Save Session)

Creating a Composite: Detailed Notes

First, start PROfit and Photoshop followed by the EvoFIT (via the desktop icons or from the Start menu in Windows).

The Initial Process to Create a Composite

In the EvoFIT package, begin a photofit session by selecting *Create ...* from the main menu item called *EvoFIT* (i.e. *EvoFIT->Create ...*). This will display a screen called the Face Palette that shows the first 18 random face shapes. Note, the working folder for this photofit session is displayed in the title bar; for example -

EvoFIT Face Palette -- Face Set 1: 1st Generation (C:_ExptAF\Ss\0003\)[Shape Only -> Screen
1/3]

Begin by selecting a hairstyle from PROfit and modify it in Photoshop if necessary. Follow the procedures overleaf.

The EvoFIT system automatically begins with the Shapes screen. This is reflected in the Face Palette's title bar -

EvoFIT Face Palette -- Face Set 1: 1st Generation (C:_ExptAF\Ss\0003\)[**Shape Only** -> Screen
1/3]

Select *at least* 6 face Shapes (by clicking on the best matches with the left-hand mouse button).

Select shapes from *at least 3 screens* of random facial shapes; more screens can be seen by selecting *More Faces* from the *Face Palette's Display* menu (i.e. *Display->More Faces*). The number of screens displayed is similarly shown in title bar -

EvoFIT Face Palette -- Face Set 1: 1st Generation (C:_ExptAF\Ss\0003\ [Shape Only -> Screen 1/3]

Switch to the *Textures* screen (*Display->Type->Textures*). Note that the title bar changes to reflect that *Textures* are being displayed -

EvoFIT Face Palette -- Face Set 1: 1st Generation (C:_ExptAF\Ss\0003\ [Texture Only -> Screen 1/3]

Select at least 6 face textures over at least 3 screens (via *Display->More Faces* as before).

Switch to the *Normal* screen (*Display->Type->Normal*). Note that the title bar changes to reflect that *both facial shapes and textures* are being displayed now; for example -

EvoFIT Face Palette -- Face Set 1: 1st Generation (C:_ExptAF\Ss\0003\ [Normal]

On the *Normal* view, faces with a *blue* border were selected for *Shape*, faces with a *green* border were selected for *Texture* and those faces with a *red* border were selected for both *Shape* and *Texture*. The faces with a red border should be better than the ones with a blue or green border.

First, select any of the faces with a blue or green border that are as good or better than the ones with a red border (by left-clicking with the mouse).

Next, select the face that is best overall (by *right*-clicking with the mouse and selecting *Best Face* from the pop-up menu).

The First Set of Offspring Faces

Press the *Make Next Generation* button (or *Evolve->Next Generation*) to generate the *first* set of *Offspring* faces. Use the rating scale (overleaf) to record the quality of the photofit when requested.

The offspring faces will be displayed as facial *shapes* initially (as for the first set of faces). These new faces are based on your previous Shape and Texture selections. Once again, select at least 6 facial shapes over at least 3 screens. One thing different for the *Offspring Shapes* is the shapes are shown with the texture from the *best face* selected in the previous generation (the very first generation of shapes used a smooth or averaged texture). This is aimed at speeding up the evolution process. It is sometimes the case that a particularly good face overall is produced via the use of the *best texture*. So, if you consider that any of these faces are as good (or preferably) better than the previous *best face*, right-click on the face and then select *Make Normal* from the pop-up menu. This will ensure that this face is present when you later display the population of faces in *Normal* view.

Once at least 6 shapes have been selected, change to the *Textures* Screen and select at least 6 facial textures over at least 3 screens (as before). In a similar way to the display of Offspring Shapes, textures are displayed with a shape from the *best face*. Once again, use the right-mouse button to save a particularly good face (for display in the *Normal* view).

As before, next switch to the *Normal* view and select any faces with a blue (Shape) or green (Texture) border that can be considered as good or better than the faces with a red (Shape & Texture) border. Select the best face with the right mouse button.

If desired, the *best face* can be improved using the *Feature Shifter*. Follow the procedure overleaf. When finished, you will return to the *Normal* display with the selected face replaced with your modifications (if you chose to save them). Remember, the hairstyle can be tweaked in Photoshop as well (in fact at any time); follow the procedure overleaf for that too.

Create another generation of faces (press the *Make Next Generation* button or select *Evolve->Next Generation* from the menu).

Continue the process of Selecting, improving and generating faces until an acceptable likeness is achieved. Follow the procedure overleaf to save this face to disk as the final EvoFIT. It is

recommended that about 4 complete generations be bred before finishing.

Note also, that it is possible to return to and/or continue from a previous screen at any time during evolution; use *Display->Previous Face Set* and *Display->next Face Set* from the *Face Palette* menu.

When finished, select *EvoFIT->Exit (Save Session)* from the *Face Palette* menu. Do NOT use *EvoFIT->Cancel* as your session will be over-written the next time an EvoFIT is created (see me if this accidentally happens).

Selection of a Hairstyle in the PROfit System

Ensure that the PROfit system is running.

Select *Edit->Modify hair & Internal Features* from the EvoFIT *Face Palette's* menu to enter the *Modify Hair* utility (for hairstyle selection and modification). Ensure the following is set (should be set by default) -

- *External Features* is selected in the *Edit* group
- *PROfit* is selected in the *Editor* group
- *Current Session* is selected in the *Source* group

Click the button marked: *Load into Editor (from Source)*. This will copy the current session background (called *ef.bmp*) to the PROfit folder.

Do the following in the PROfit system (refer to the PROfit (CD-FIT) manual for more details): *Load* the file **ef.bmp** into PROfit as a *face* in *Feature Editing*. When the *Picture Definition* window appears (to enable the user to select the edges of a feature to import), ensure that the selection box fits just *outside* the left and right vertical edges of the EvoFIT image. Next, move the *FACE* acetate so that it is immediately beneath *HAIR* (press the *Config* button and then select *Layer Order*). Finally, select the icon for *hair* (in the face in the top right-hand corner of the screen). You can now ready select and manipulate a hairstyle in PROfit.

When a hairstyle has been found, save the composite to disk: press the *Commands* button, then *Save PROfit*. The *file name* must be called *ef* and the *File Format* must be **BMP**.

Return to EvoFIT's *Modify Hair* utility. Press the *Import Photofit* button and set the red guidelines so that they are *flush* with the border of the EvoFIT image. Set all 4 lines such that none of the white background is seen (don't cut into the EvoFIT image itself). Press the *Import* button and notice the *Face Viewer* window (normally on the left-hand side of the screen) update appropriately. Re-import the image if it is not acceptable (there should not be any *extra* black border added nor should the EvoFIT border be cropped in any way). Press the *Close* button to exit from the *Import Photofit* utility. Note, importing only updates the temporary or working background (and therefore does not change the background of population faces).

The hairstyles in the *Face Palette* can now be updated by clicking *Apply* in the *Update Population* group (ensure that *External Features* are also selected in the *Update Population* group before using *Apply*) of the *Modify Hair* utility.

Click the *Exit* button in the *Modify Hair* window. The utility returns to the *Face Palette* and the population faces are then updated.

This utility can be re-run at anytime. As the hairstyle has already been loaded in to PROfit, it is NOT necessary to re-load it (i.e. Don't press *Load into Editor*, just save changes in PROfit, *Import* the photofit and then *Apply* changes).

Modification of a Hairstyle in Photoshop

Ensure that Photoshop is running.

Select *Edit->Modify hair & Internal Features* from the EvoFIT *Face Palette's* menu to enter the *Modify Hair* utility (for hairstyle selection and modification). Ensure the following is set -

- *External Features* is selected in the *Edit* group
- *Photoshop* is selected in the *Editor* group
- *Current Session* is selected in the *Source* group

Click the button marked: *Load into Editor (from Source)*. This will load the current session's hairstyle (called *ef.bmp*) directly in to Photoshop.

Remember to switch the *Mode* to *RGB Color* in Photoshop for best editing (*Image->Mode->RGB Color*). When finished, switch the image mode back to greyscale (*Image->Mode->Greyscale*). It is important that the image is not resized. Save the image (*ef.bmp*) to disk.

Return to EvoFIT's *Modify Hair* utility and click *Apply* in the *Update Population* group (ensure that *External Features* are selected in the *Update Population* group before using *Apply* - as before).

Click the *Exit* button in the *Modify Hair* window. The utility returns to the *Face Palette* and the population faces are then updated.

Moving Facial Features Around with the Feature Shifter

The *Feature Shifter* utility allows facial features to be moved and resized. It is activated from the *Normal* display mode by right-clicking an image in the *Face Palette* and selecting *Feature Shifter* from the pop-up menu. The *Face Palette* will be hidden and two windows will appear: an *Image Viewer* window and an *Image Parameter* window.

Modes of Operation

There are 2 modes that the *Feature Shifter* can operate: a *holistic* morph and a *free* morph. Selection is made between the modes in *Morphing* group (the top left-hand group box in the *Image Parameter* window). Both types will move features specified by selected pixels (e.g. mouth points). Whereas a *free* morph will perform a facial-feature morph merely as coordinate points change, the *holistic* morph does a best fit in the holistic shape model first (i.e. before the feature morph). This means that the holistic morph keeps the face as a holistic representation (which we believe to be a very good thing). The problem comes if the target trying to be created is not well represented in the shape model (e.g. the eyebrows have been trimmed or "tampered" in some way). In this case, trying to get short eyebrows is likely to "over-stretch" the shape model, and a *free* morph is preferable. Overall, it is probably best to use the *Feature Shifter* in *holistic* mode until towards the end of the photofit session, and then switch to *free* morph for final tweaks. Note, that the *Feature Shifter* automatically starts in the *holistic* mode.

Moving Features Around

To move facial features, first select the desired feature from the *Image Viewer* menu; e.g. *Edit->Mouth->All*. The features to be moved are highlighted by coordinate points shown in *red*. Then, use the control buttons under the *Image Viewer* menu to manipulate the face. Note that the *step size* of any change can be altered via the edit box in the top left-hand corner of the *Image Properties* window. Note also, that only selected coordinates (those that appear in *red*) will result in a morph. Specifying a feature to move, for example *Edit->Mouth->All*, just activates a group of coordinates quickly. You can select individual coordinates to be moved by clicking with the mouse. Coordinates can be returned to their inactivated state (green) by selecting *Edit->Clear* from the *Image Viewer* menu.

Working with Temporary Backups

Just like it is useful to keep backups of text documents in case of mistakes, this same idea is available for faces in the *Feature Shifter*. It is a good idea to create a temporary or working copy after *each* feature change has been carried out. This will enable you to return to the last stored face if an undesirable change has occurred.

A particular face can be stored and recalled via the controls in the *Temporary Backups* group box in the top right hand corner of the *Image Properties* window. The *location* number indicates where a current image will be stored and retrieved. Click *Store* to save a

copy of the current face in the location shown and *Recall* to retrieve it. All changes can be undone by the *Return to start* button.

Size of Morphed Changes

The *step size* of any morphing change is set via the *Morphing* group box in the top left-hand corner of the *Image Properties* window. In the *free* morph mode, coordinate movements will be equal to the number specified for the step size. This will not necessarily be the case for a *holistic* morph. In this mode, a best fit is carried out in the Shape Model before a morph is performed. The *actual* movement obtained now depends largely upon the *number* of coordinates selected; *fewer* coordinates selected will require *larger* step sizes. For instance when moving the nostrils up, a group of only 8 coordinates, a step size of at least 2 or 3 will be required.

Saving

When satisfied with changes, use *EvoFIT->Exit (Replace Face in Population)*. Changes can be discarded by *EvoFIT->Cancel (Discard Changes)*.

Saving a Face as an EvoFIT

A face must be saved as a photofit from within the *Face Palette*. To do this, make sure that the EvoFIT system is in the *Normal* display mode with the desired face selected. Select *EvoFIT->Save As* from the *Face Palette's* menu. Name the file sensibly (e.g. **EvoFIT1.bmp**); the correct path for the EvoFIT should already be selected (in a folder under the current session folder). If subsequent modifications are to be made (for example in Photoshop), it is best to work on a copy (e.g. **EvoFIT2.bmp**) rather than the image just saved (in case of editing "accidents"); e.g. make a copy of **EvoFIT1.bmp** and rename it as **EvoFIT2.bmp**.

Once this has been saved successfully, exit the *Face Palette* to save the session before creating another photofit: select *EvoFIT->Exit (Save Session)* from the *Face Palette* menu. Do NOT select *EvoFIT->Cancel* as your session will be over-written the next time an EvoFIT is created (see me if this accidentally happens); only use *EvoFIT->Cancel* if session changes are to be lost intentionally.

Printing an EvoFIT

A photofit must first be saved to disk before printing (see the section on Saving a Face as an EvoFIT). Printing can be done from within the *Face Palette* or from the main EvoFIT application window (i.e. the window that appears when you first start the EvoFIT program). In either case, select *EvoFIT->Print ...*

Click on the first *Select* button (in the top right-hand corner of the Print dialog window) adjacent to the box for *EvoFIT*. A dialog window appears that allows the previously saved EvoFIT to be selected. Select the EvoFIT (e.g. **EvoFIT1.bmp**) and click *Open*. You will see the file's path appear to the left of the *Select* button. Click the *Preview* button and a *Print Preview* window should appear (occasionally, this window appears behind the others; use the windows task bar (at the bottom of the screen) to bring it to the foreground). Select *File->Print* from the *Print Preview* window to print.

When finished printing, close the *Print Preview* window and then the EvoFIT *Print* dialog window with the *Exit* button.

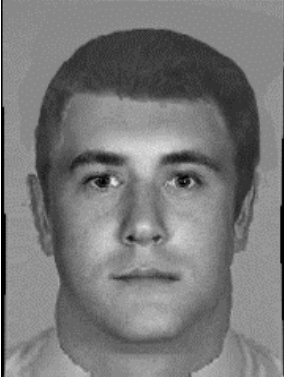
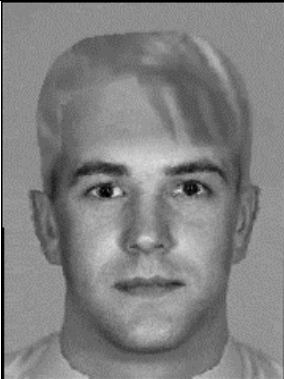
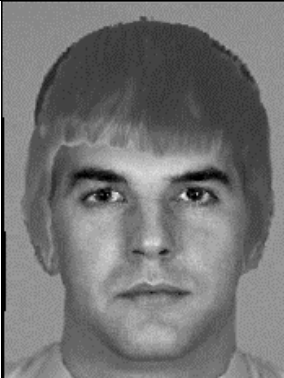
Rating Scale

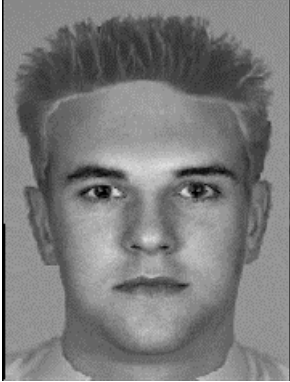
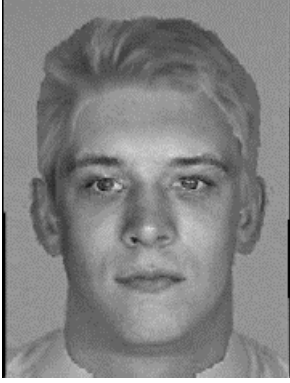
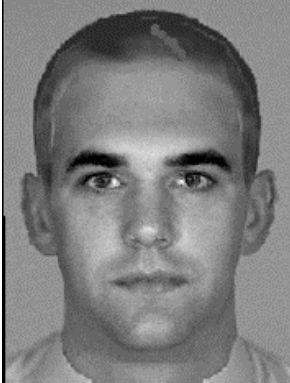
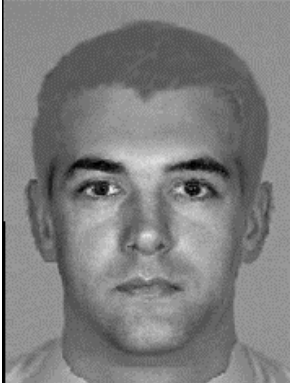
Use the following scale to rate the similarity of the *best face* to the target when requested (when generating a new population of faces) -

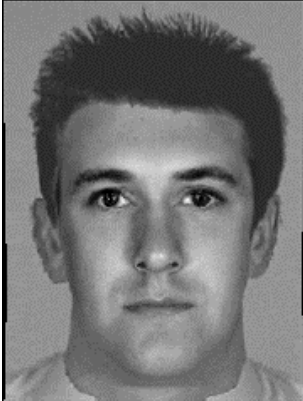
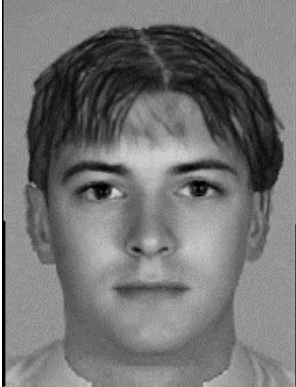
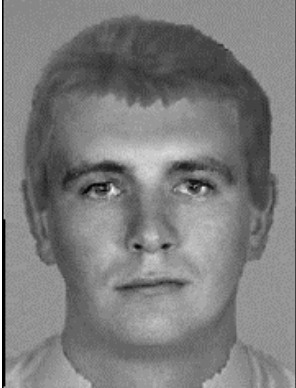
- | | |
|--------|----------------------------------|
| 1 | Very poor likeness between faces |
| 2 or 3 | Few similarities |
| 4 or 5 | Some similarities |
| 6 or 7 | Many similarities |
| 8 or 9 | Faces could be easily confused |
| 10 | Faces are identical |

Appendix G: EvoFITs created in Experiment 11

The following are the EvoFITs evolved in Experiment 11.

Celebrity	Target
Craig Phillips	 A grayscale portrait of a young man with short, dark hair, looking directly at the camera with a neutral expression.
David Beckham	 A grayscale portrait of a young man with short, light-colored hair, looking directly at the camera with a neutral expression.
Noel Gallagher	 A grayscale portrait of a young man with short, dark hair and a fringe, looking directly at the camera with a neutral expression.

<p>Leonardo DiCaprio</p>	
<p>Matt Damon</p>	
<p>Robbie Williams</p>	
<p>Michael Owen</p>	

<p>David Schwimmer</p>	
<p>Stephen Gately</p>	
<p>Tim Henman.</p>	

Appendix H: Flowcharts for Face Generation and EvoFIT Operation

The following flowcharts summarize EvoFIT face generation and use for the full system evaluated Experiment 10 and Experiment 11. Flowcharts consider firstly the initial generation, with random faces, and then subsequent generations (generation 2 and over). The following abbreviations are used:

FNP Facial Normal Palette

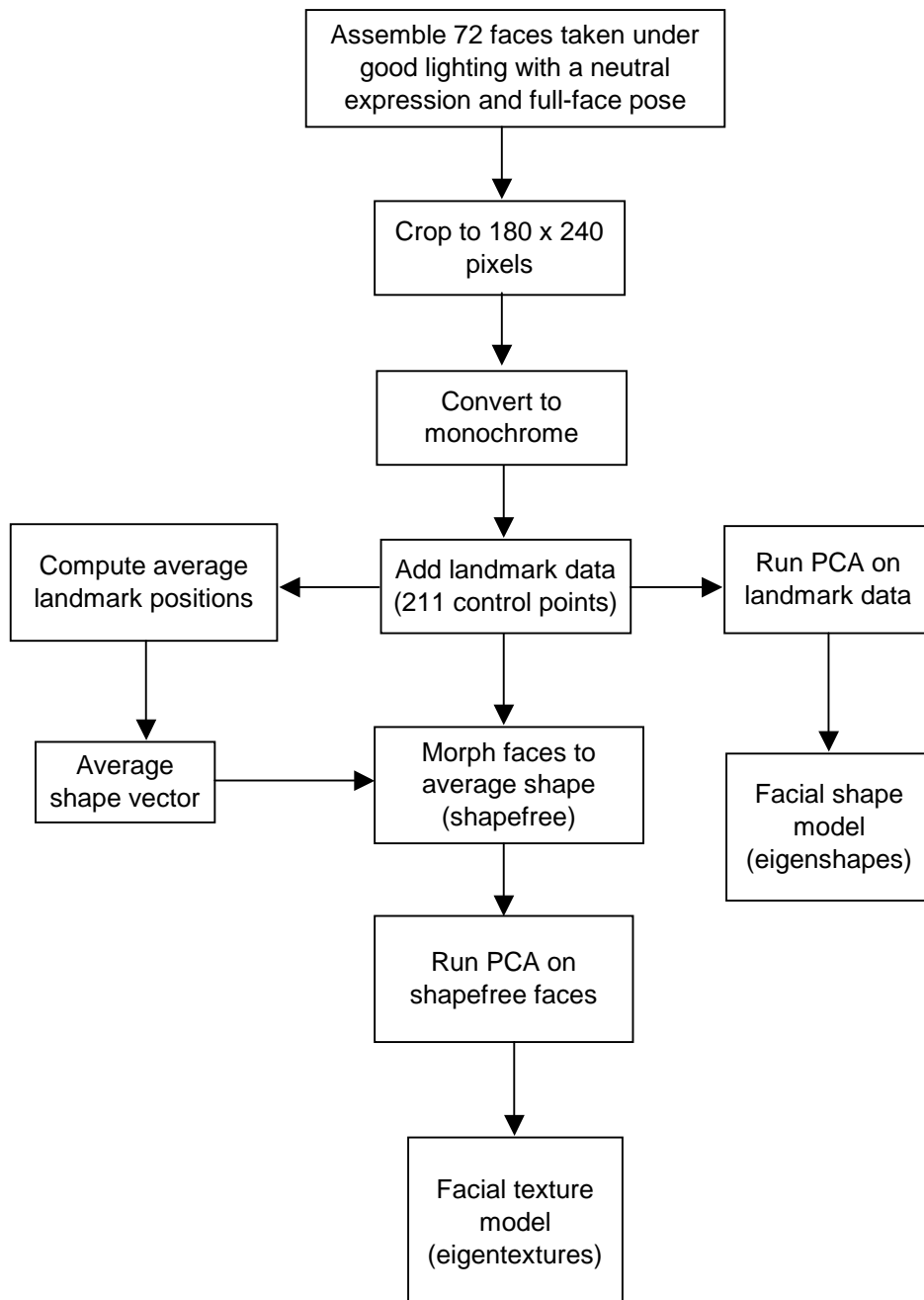
FSP Facial Shape Palette

FSPBT Facial Shape Palette with the Best Texture (from the previous generation)

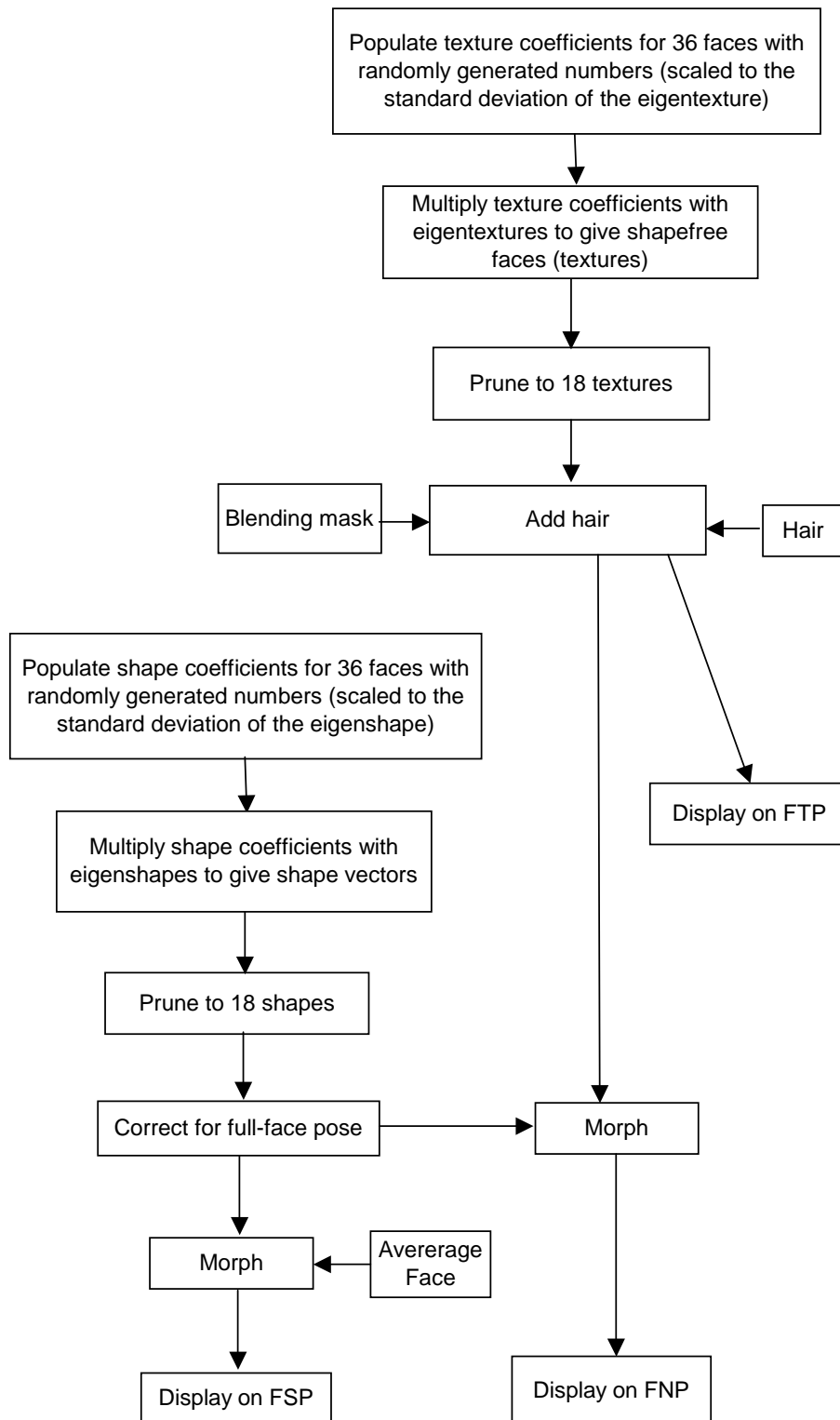
FTP Facial Texture Palette

FTPBS Facial Texture Palette with the Best Shape (from the previous generation)

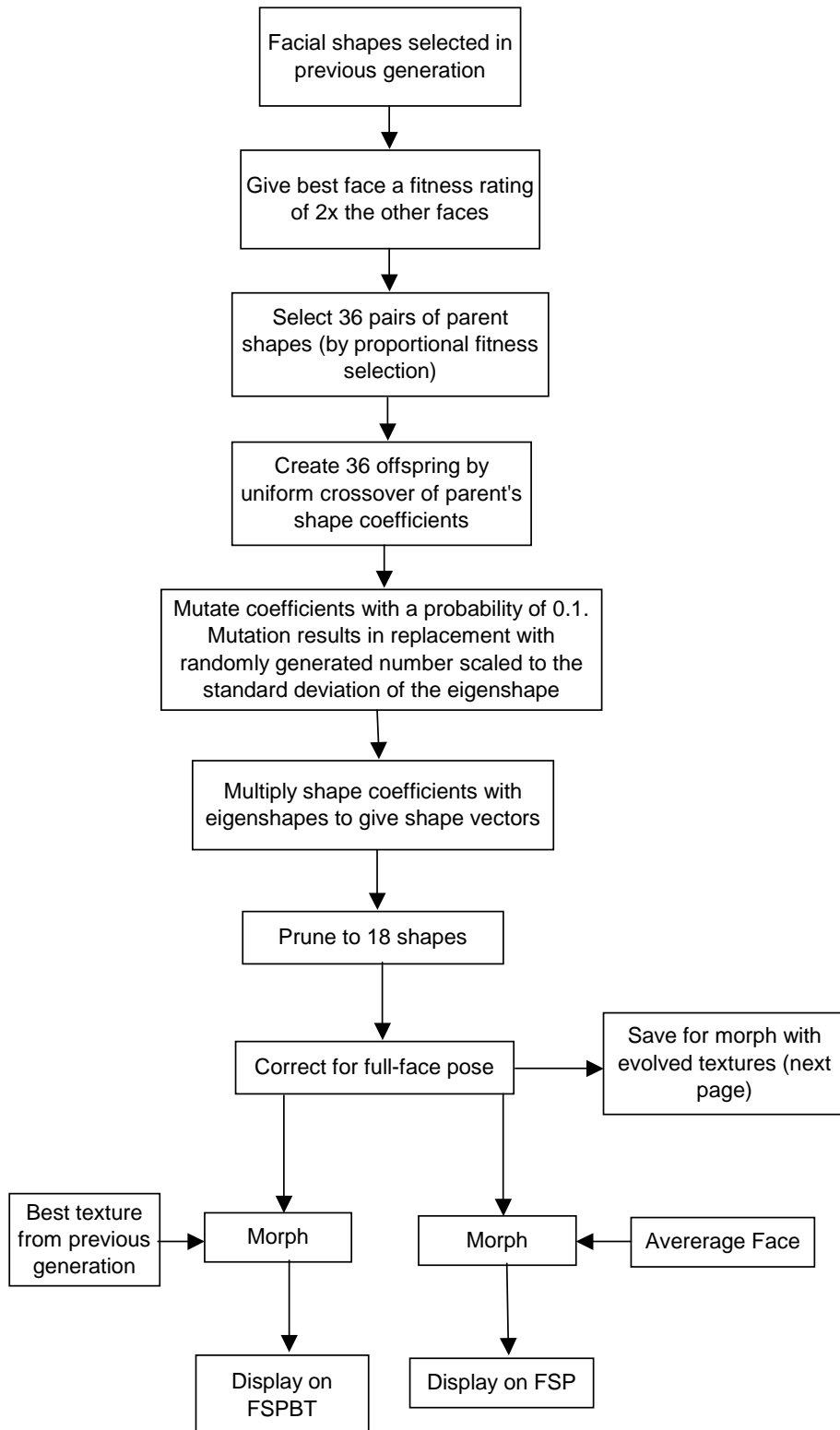
Setting up the Face Model



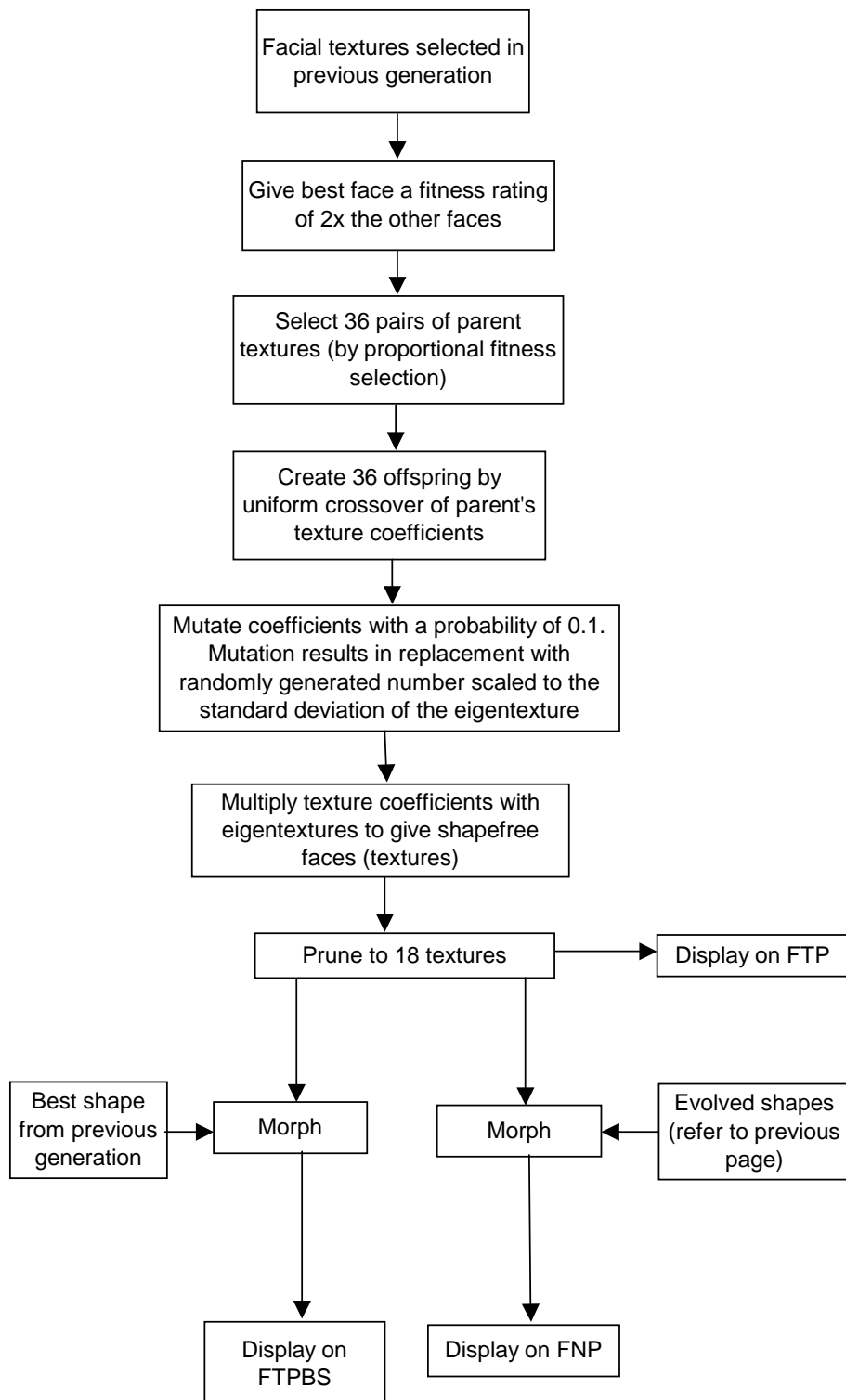
Creating Population Faces: First Generation



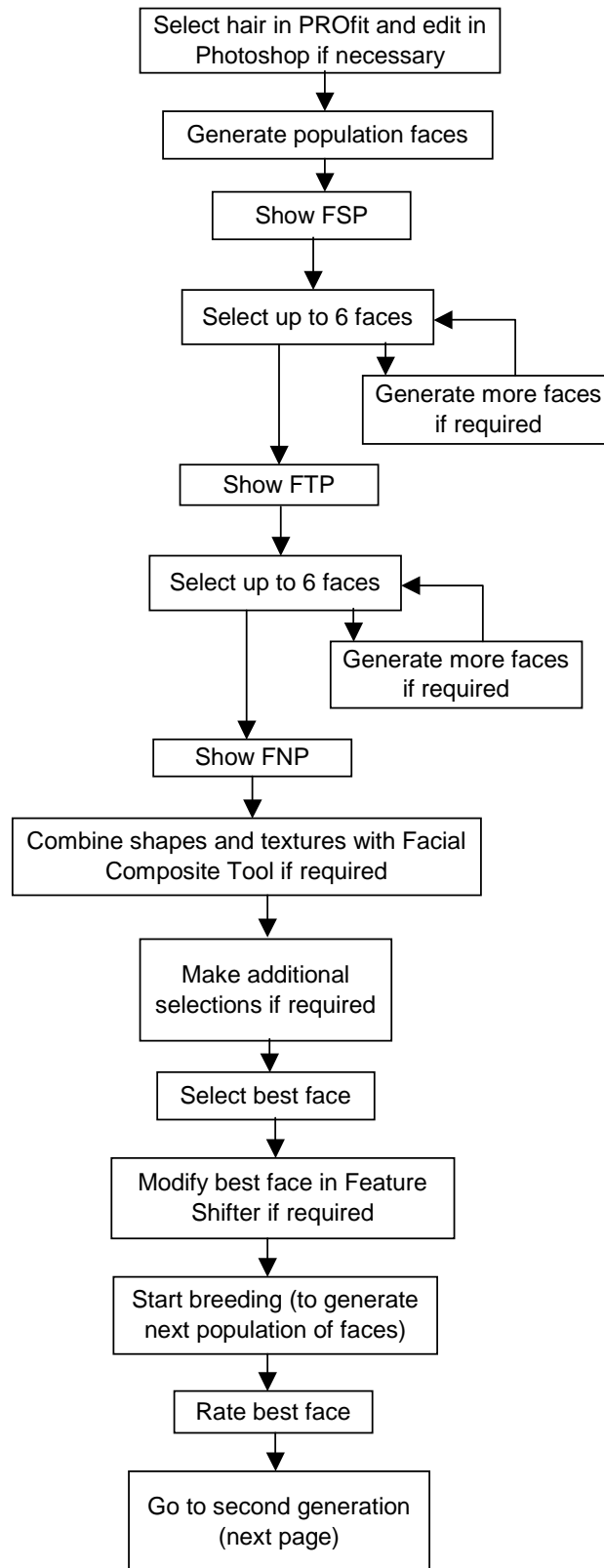
Creating Population Faces I: Generation Two and Thereafter



Creating Population Faces II: Generation Two and Thereafter



Procedure for Witness: First Generation



Procedure for Witness: Generation Two and Thereafter

